Examiner judgments of collocational proficiency in L2 English learners' writing

by

Benjamin Naismith

Bachelor of Music, Thompson Rivers University, 2005

Master of Science, Aston University, 2016

Master of Arts, University of Pittsburgh, 2019

Submitted to the Graduate Faculty of the

Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Benjamin Naismith

It was defended on

February 11, 2022

and approved by

Matthew Kanwit, Associate Professor, Department of Linguistics

Melinda Fricke, Assistant Professor, Department of Linguistics

Ute Römer, Associate Professor, Department of Applied Linguistics and ESL, Georgia State University

Dissertation Director: Alan Juffs, Professor, Department of Linguistics

Examiner judgments of collocational proficiency in L2 English learners' writing

Benjamin Naismith, PhD

University of Pittsburgh, 2022

This study investigates how aspects of collocational proficiency affect the ratings that expert examiners give to second language (L2) English learner essays. Lexical proficiency is a multi-faceted phenomenon and certain aspects of it are particularly impactful on human judgements, including lexical sophistication and accuracy. However, the importance of proficiency with formulaic sequences (FSs), like collocations, has received less attention than proficiency with single words, despite FSs' essential role in language production. In addition, previous comparison studies have used a small number of raters with varying levels of assessment expertise, assessing texts of varying length and topic.

In addressing these issues, this study uses a predominantly quantitative, experimental approach comprised of two stages. First, a small set of three texts of different proficiency levels were created based on model IELTS Task 2 essays, controlling for topic and length. From these texts, a set of 30 versions were produced, manipulating specific collocational features related to sophistication and accuracy. Second, IELTS examiners (n = 47) rated the texts and provided rationales for their choices. From these data, many-faceted Rasch models were used to obtain expected scores, and linear regression models were used to determine which aspects of collocational proficiency best predicted the experts' ratings.

The findings reveal that increases in lexical sophistication significantly and positively impacted the experts' ratings. Post-hoc analyses demonstrated that the categories of high sophistication and mid sophistication differed significantly from low sophistication. However, mid

sophistication was not significantly different from high sophistication. When these 'advanced' words were used as part of collocations, they then provided a small but significant additional boost to ratings. Notably, there was no significant effect for increased collocational accuracy. In conjunction, these findings indicate that 1) sophistication is perhaps best viewed on a spectrum rather than categorically, 2) there is an additional increase to ratings if learners use advanced lexis as part of collocations, and 3) there is a potential baseline in terms of gravity and frequency of collocation errors below which ratings are not significantly affected. The implications for these findings are therefore discussed in relation to written language assessment and L2 vocabulary pedagogy.

Table of contents

List of tables

List of figures

Acknowledgments

For those of you who know me, it will come as no surprise that I put off writing this section until the very last moment. Outpourings of earnest and effusive praise do not come naturally to me, but in this case that is exactly what is merited. As anyone who has completed a PhD knows, it is a collaborative effort, even if the title page would suggest otherwise. And so with that awkward preface, let the gushing begin…

First off, my deepest gratitude is for my advisor, Dr. Alan Juffs, without whom none of this would have been possible. From the first tentative email I sent to him expressing interest in Pitt, Alan has been unfailingly supportive. Over the last five years, he has taken on the roles of advisor, mentor, professor, advocate, colleague, co-author, and friend. If I should someday find myself in a similar position advising graduate students, I will certainly aim to emulate his model.

Most graduate students are lucky if they find one advisor to admire, but in my case I somehow ended up with two. From day one at Pitt, my second unofficial advisor has been Dr. Na-Rae Han, performing many of those same roles I have listed for Alan. Under her guidance, I am leaving Pitt with a skillset I never imagined I would possess: a proficiency in computational and corpus linguistics that allows me to answer the research questions that have always fascinated me. For inviting me to be part of the Pitt ELI Data Mining Group, I will always be grateful.

I have also interacted and learned from many other excellent professors during my studies, too numerous to list here. However, I would like to acknowledge the incredible support from my other dissertation committee members. Dr. Matthew Kanwit has never ceased to be an incredible resource and model, supporting all aspects of my professional development in the classroom and in my research. Likewise, Dr. Melinda Fricke has always graciously given her time over the years to help with statistical consulting regardless of the project. Finally, there is Dr. Ute Römer, one of

my linguistics idols, who graciously agreed to have coffee with me at a conference. This small act of kindness somehow led to the good fortune of having her involvement in this project.

For this research, I also owe a giant debt of gratitude to my network of friends and colleagues around the world who gave up their time to freely publicize and participate in this study. The global English language teaching community is truly amazing. Likewise, I am grateful to all of the students who contributed data to the PELIC corpus and to the teachers and staff at the Pitt English Language Institute, especially Dawn McCormick for her support over the years. And of course, a giant thankyou goes out to all of the graduate students who I am proud to call my friends. You are too many to name, but you know who you are.

I am equally grateful to everyone in the Pittsburgh community who has welcomed my family into this city. I would be amiss not to mention Dorolyn Smith for providing us with our home, and the Western Pennsylvania School for the Deaf for their care of my daughter. I have also received invaluable funding from a variety of sources which has been critical to our wellbeing. I greatly appreciate this financial support, most notably from the University of Pittsburgh, the Social Sciences and Humanities Research Council of Canada, Duolingo, and the Humanities Engage project.

Finally, I have saved my most important thankyous for last. To my family in Canada, your unwavering love and support means the world to me. To my daughter Mia, you are the reason that I have worked so hard to be the best version of myself that I can be. I hope one day that you'll be as proud of me as I am of you. And to my wife, Andrea, I literally could not have done this without you. We have spent over a decade travelling the world together, and you have supported me in every way imaginable – this degree is as much yours as it is mine. I can't wait for our next adventure (though perhaps no PhDs for a while). I love you all.

1.0 Introduction

In the field of second language (L2) assessment, there is a long history of assessing extended production of writing (Lenko-Szymanska, 2019). At present, the importance of such assessment has never been greater as it is a major component of high-stakes international proficiency exams like the International English Language Testing Service (IELTS) and the Test of English as a Foreign Language (TOEFL). The results of such tests can be incredibly consequential for candidates in contemporary society (Crossley et al., 2014; McNamara et al., 2019), determining in part whether they are able to study or work abroad (IELTS, 2019). Beyond the testing context, academic writing proficiency is also a key measure of academic success more generally (Kellogg & Raulerson, 2007; Leki & Carson, 1994). However, acquiring this proficiency in English is a challenge regardless of first language (L1; Kyle & Crossley, 2016) because the requisite academic language extends far beyond that which is used in basic interpersonal communication (Cummins, 2003; Simpson-Vlach & Ellis, 2010). It is therefore imperative for stakeholders in English Language Teaching (ELT), including teachers, material developers, and curriculum designers, to carefully consider how to best support learners in achieving the required level of academic writing proficiency.

Lexical features are particularly important in this regard, central to the process of language acquisition (Cobb & Horst, 2015; Schmitt, 2010) and developing writing proficiency (Dabbagh & Janebi Enayat, 2019; Kyle & Crossley, 2016; Lee, 2003; Levitzky-Aviad & Laufer, 2013; Ruegg et al., 2011). As with writing, vocabulary is a key predictor of academic achievement (Qian & Lin, 2020; Roche & Harrington, 2013, 2014) and learners themselves are often aware that vocabulary is a key element of language learning and writing (Ellis, 1995; James, 1998; Leki & Carson, 1994;

Polio & Glew, 1996). It is therefore unsurprising that there has been a proliferation of vocabulary studies in the last 30 years (Bulté et al., 2008; Daller et al., 2007; Gyllstad, 2013; Nation, 2011), especially with the advances in computational linguistics. As a result, many aspects of lexical proficiency have been defined and quantified, though determining precisely which lexical features to prioritize remains a matter of debate.

One promising avenue of investigation is the importance of formulaic sequences (FSs) in writing proficiency (Monteiro et al., 2020; Wray, 2000), e.g., collocations (to be operationalized in Section 3.2). Traditionally, single words were the basis for most lexical indices, but more recently studies have focused on collocations and their impact on readers (Crossley et al., 2012; Durrant, 2019; Granger & Bestgen, 2014). Within this context, this dissertation investigates examiner judgments of collocational proficiency in L2 English learners' writing. To do so, statistical collocational features and expert human ratings of learner essays were compared. In contrast to previous studies, a small, carefully constructed set of texts were rated by a large pool of standardized expert raters. From the data, the research determined which aspects of collocational proficiency best predict the ratings of assessment experts, focusing on collocational sophistication and accuracy. In addition, qualitative data were collected about the reasons for the raters' scores. These secondary data are intended to better understand which lexical features are most salient and attended to by expert raters[1]. The motivation for this research is to build on the existing body of research in this field by addressing the following research gaps:

1. consideration of collocational features and their impact on expert raters' judgments
2. consideration of collocational accuracy

[1] The study is not a formal mixed methods study in the sense of Creswell and Plano Clark (2011).

3. the distinction between mid- and low-frequency lexis in considering lexical/collocational sophistication
4. limited studies using a large number of expert raters
5. limited studies using texts with identical lengths and prompts
6. limited studies adequately accounting for inter-rater unreliability

The structure of the dissertation is as follows: Chapter 2 reviews the broad concept of *lexical proficiency* and dimensions relevant to this study. Chapter 3 then narrows the focus to one type of lexical unit, collocations, in a discussion of *collocational proficiency.* Chapter 4 is the final chapter of the literature review, describing studies which compare human ratings and statistical measures of texts. Chapter 5 presents a preliminary study, a description and validation of a process for normalizing essays of different lengths while maintaining their key characteristics. These essays are then used in Chapter 6 as the instruments for the main study, described above. This chapter includes the methodology and findings of the main study. Chapter 7 concludes the dissertation by presenting a discussion of the findings and the implications of this research.

2.0 Lexical proficiency

Before analyzing *collocational proficiency*, it is necessary to first consider the broader concept of *lexical proficiency*. In this chapter, this term is defined, and we consider how lexical knowledge, especially productive knowledge, is realized in learner writing. First, the dimension of lexical breadth is discussed, focusing on the importance of lexical sophistication. Methodological considerations are also presented with respect to frequency bands and counting units. Second, lexical accuracy is considered, highlighting its impact on readers and the ways in which it is typically measured.

Lexical proficiency is a critical element of L2 language learning and impacts not only writing, but all other systems and skills, including reading (Laufer, 1992; Roche & Harrington, 2014), listening (Bonk, 2000; Stæhr, 2008), and speaking (Milton, 2013; Milton et al., 2010). And yet, although the term 'lexical proficiency' is often used in vocabulary research, it has remained difficult to define (Crossley & Skalicky, 2019), at least in a way which achieves broad consensus. This confusion is in part due to lexical terminology which is often used inconsistently across studies (Bulté et al., 2008), with reference to overlapping abstract lexical concepts (Elgort & Siyanova-Chanturia, 2021): 'lexical proficiency' is used interchangeably with 'lexical competence', which in turn is substituted for 'vocabulary knowledge' or 'lexical knowledge'. In this paper, 'lexical competence', 'lexical knowledge', and 'vocabulary knowledge' will be considered synonymous,[2] and the term 'lexical knowledge' will be used whenever possible for

[2] 'Lexical competence' is sometimes logically described as encompassing both word knowledge and ability (Lenko-Szymanska, 2019). However, this distinction is often not adhered to across the research in this area.

consistency. However, I will disambiguate 'lexical knowledge' from 'lexical proficiency': moving forward, proficiency refers to "an ability to apply both declarative and procedural lexical knowledge in real language use" (Lenko-Szymanska, 2019, p. 39). In other words, proficiency is the manifestation of lexical knowledge (Bulté et al., 2008). Furthermore, in this study, I take the position that when we observe the output of L2 learners, we assume that their production reflects their lexical proficiency, which in turn is thought to tap into their lexical knowledge (Crossley & Skalicky, 2019; Kim et al., 2018).

2.1 Lexical knowledge

It is commonly asserted that vocabulary knowledge is 'multi-faceted' (e.g., Clenton & Booth, 2020; Daller et al., 2007; Qian & Lin, 2020), encompassing numerous aspects (Elgort & Siyanova-Chanturia, 2021; González-Fernández & Schmitt, 2019). The most widespread model of these in applied linguistics is that of Daller et al. (2007, p. 8) which depicts lexical space as being three-dimensional, comprised of *breadth*, *depth*, and *fluency* (see Figure 1).[3] Essentially, as first described by Anderson and Freebody (1981), breadth refers to the number of words a person knows, and depth describes how well the words are known. Of the two, depth is widely acknowledged as being more difficult to measure, regarded by Schmitt (2014, p. 920) as "the wooliest, least definable, and least operationalizable construct in the entirety of cognitive science." Typically, therefore, individual components of lexical depth are studied in isolation (González-

[3] For an overview of other theoretical models of the lexicon, see Juffs (2009) including perspectives from generative SLA (e.g., Jackendoff, 2002) and psycholinguistics (e.g., Kroll & De Groot, 1997).

Fernández & Schmitt, 2019), for example, derivational knowledge or collocational knowledge (though see Chen & Truscott, 2010; Naismith & Juffs, 2021; Schmitt, 1999; Webb, 2007 for examples of multi-component lexical depth studies).

The third category, 'fluency' (sometimes referred to as 'access'), is a later addition to the model and refers to the ability of a person to access their stored lexical items. In writing, fluency is difficult to measure if only the final text is considered, though it may be simplistically operationalized as the number of words produced (Lenko-Szymanska, 2019). Lexical fluency will not be described further, though it is undoubtedly a critical element of any complete model of lexical proficiency.

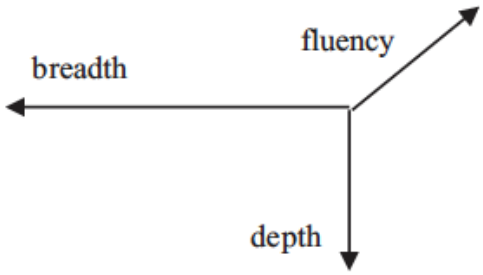

Figure 1 The model of lexical space (Daller et al., 2007, p. 8)

Returning to breadth and depth, separating the two concepts is no easy task, leading Schmitt (2014) to question the validity of this distinction. Teasing apart breadth and depth is a challenge because there is always a high correlation between the two; a person who knows more words typically also has deeper knowledge of words (González-Fernández & Schmitt, 2019; Qian & Lin, 2020). Nevertheless, recent analyses using Structural Equation Modeling (SEM) have substantiated the argument for keeping them as distinct constructs, based on their unique contributions to overall models of lexical knowledge (González-Fernández & Schmitt, 2019; Koizumi & In'nami, 2020; Vafaee & Suzuki, 2020). For the breadth/depth dichotomy, there is also

perceived widespread ecological validity in that this dichotomy can be frequently observed naturally outside of research settings. For example, the vocabulary descriptors used in the Common European Framework of Reference (CEFR; Council of Europe, 2001) extensively use the terms 'range' and 'control', which equate to 'breadth' and 'depth' respectively (Milton, 2013).

2.1.1 Productive and receptive knowledge

One final way of subdividing lexical knowledge is through the contrast of *productive knowledge* and *receptive knowledge*, i.e., the ability to produce a word versus the ability to comprehend a word (Laufer et al., 2004). These two types of knowledge are explicitly linked to the four skills, with productive knowledge tied to speaking and writing, and receptive knowledge tied to reading and listening. This distinction is an important one as vocabulary knowledge for a speaker may be inconsistent across the four skills and should therefore be treated differently depending on which skill is under investigation (Clenton & Booth, 2020).[4] For example, Stæhr (2008) found that although vocabulary size correlated with reading, writing, and listening, the correlation differed for each skill, with vocabulary and reading correlating most strongly. In general, learners tend to know more items receptively than productively (Lee, 2003; Milton, 2009), especially for low-frequency words (Schmitt & Meara, 1997; Waring, 1997). This discrepancy can be attributed to the more complicated nature of productive knowledge (Nation, 2016) which requires a series of mental processes to be carried out before production can take place (Kormos,

[4] Within productive knowledge, oral and written knowledge is also distinct, with advanced learners developing their written vocabulary at a greater rate than their spoken (Milton et al., 2010), though these differences will not be explored in this paper.

2006; Levelt, 1989, 2001). Nevertheless, it would be incorrect to view receptive and productive knowledge as being completely independent, with various studies showing a positive correlation between receptive vocabulary knowledge and productive writing proficiency (e.g., Koda, 1993; Schoonen et al., 2011).

To better account for the types of productive/receptive knowledge, Laufer et al. (2004) proposed a more fine-grained two-way distinction of *active/passive* and *recall/recognition.* In this conceptualization, *active knowledge* signifies that a user can retrieve a word, and *passive knowledge* signifies that a user can supply a meaning for a word presented to them. *Recall knowledge* is when a user can recall either the form or meaning of a word, whereas *recognition knowledge* suggests that a user is able to do so, but only if presented with a set of options. The importance of this taxonomy is that the four knowledge types form a clear implication scale from easiest to hardest to acquire (Figure 2), with active knowledge more challenging to acquire than passive knowledge, and recall knowledge more challenging to acquire than recognition knowledge (González-Fernández & Schmitt, 2019; Laufer et al., 2004; Laufer & Goldstein, 2004; Levitzky-Aviad & Laufer, 2013):

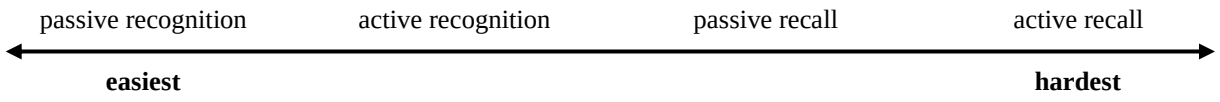

Figure 2 Types of vocabulary knowledge

For the purposes of this study, we will not be exploring these differences further. However, it is pertinent to note that using lexis in essay writing requires active recall, i.e., the strongest form of lexical knowledge and the most difficult to acquire.

We now examine aspects of productive lexical breadth and depth relevant to the study as they are realized in L2 learner writing.

2.2 Lexical breadth

Two common operationalizations of lexical breadth are *lexical diversity* (also known as *lexical variety*) and *lexical sophistication.* Both operationalizations are often based on the number of word types produced in a text. Lexis 'lends itself' to this type of statistical analysis (Lenko-Szymanska, 2019) because surface forms are readily analyzable. These efforts have been aided by the growing number of freely available computational tools (Kyle & Crossley, 2015) including Coh-Metrix (Graesser et al., 2004), Lexical Complexity Analyzer (Lu, 2010), TAALES (Kyle & Crossley, 2015), and VocabProfile (Cobb; Heatley et al., 2002). In addition, open-access NLP packages using the Python programming language have gained in popularity, e.g., the Natural Language Processing Toolkit (NLTK; Bird et al., 2009) and spaCy (Honnibal & Montani, 2017). Each of these tools and indices have much in their favour as they seek to capture different aspects of lexical proficiency. Of course, there is no one perfect metric for quantifying concepts as broad as 'lexical sophistication' or 'lexical diversity', which is the very reason such a proliferation of different indices exists. In this study, the focus is on sophistication. However, because measures of sophistication are often built on measures of diversity, we must first briefly consider diversity. It should also be noted that although diversity and sophistication are related, they are distinct constructs which do not necessarily correlate in learners' written production (Lu, 2012).

Underpinning most diversity and sophistication measures is frequency, and with good reason. The frequency of input a learner receives affects how language is processed across systems and skills, including for lexis, syntax, phonology, and reading (Ellis, 2002, 2004). Usage-based theories account for the importance of frequency by positing that language acquisition is exemplar-based (Ellis, 2002, 2004); learners acquire language through exposure to linguistic input from which they can then induce patterns using general cognitive mechanisms of learning (Ellis, 2004; Ellis & Wulff, 2015). Frequency has also historically been an important consideration for choosing which lexical items to teach in L2 classrooms (Schmitt & Schmitt, 2014; Vilkaitė-Lozdienė & Schmitt, 2020). Of course, frequency is not the only factor affecting vocabulary learning, with other factors such as contingency, recency, context, concreteness, and imageability playing a role (Crossley et al., 2019; Ellis & Wulff, 2015; Martin & Tokowicz, 2020), though these will not be addressed in this study.

2.2.1 Lexical diversity

At their core, lexical diversity metrics are typically calculated by counting the percentage of unique words in a text (e.g., Jarvis, 2013a; Malvern et al., 2004; Vögelin et al., 2019). That is, they are measures of productive vocabulary range (McCarthy, 2005). As a result, lexical diversity measures are considered 'intrinsic' (Meara & Bell, 2001) or 'text-internal' in that no reference to external data is needed for them to be calculated. The most basic diversity measure upon which many others are based is the type-token-ratio (TTR). TTR is calculated by simply dividing the total number of types in a text (i.e., the number of unique words) by the total number of tokens (i.e., the total number of words, including repeated words) (Cobb & Horst, 2015). Although practical, TTR has been justly criticized based on its sensitivity to text length: as the length of a

text increases, TTR will drop due to the inevitable repetition of function words (Cobb & Horst, 2015; Jarvis, 2013b; van Hout & Vermeer, 2007). A more detailed discussion of counting units will be presented in Section 2.2.4.

With the continued creation of vocabulary measures (Bulté et al., 2008), other more nuanced improvements on TTR have become common. Two in particular have become standard in the field and have been thoroughly reviewed and validated. The first is vocD, known also as D (Malvern et al., 2004; Malvern & Richards, 1997), which calculates TTR from a number of random samples then fits a curve and reports a parameter value. The second is the Measure of Textual, Lexical Diversity (MTLD; Crossley et al., 2009; McCarthy, 2005; McCarthy & Jarvis, 2007) which uses a complex sequential analysis of samples to generate a score based on the TTR in those samples. Taking a different approach to discussing lexical diversity, Monteiro et al. (2020) argue for the value of contextual diversity, i.e., "the number of unique contexts in which linguistic items appear" (p. 4). Otherwise known as 'range' or 'dispersion', contextual diversity provides an avenue to gain valuable insights in future research regarding another aspect of lexical breadth. To its detriment, contextual diversity is not text-internal, requiring range information from external corpora. As a result, contextual diversity might better be considered an aspect of sophistication (see next section), which is how it is categorized in Kyle and Crossley (2015).

As used in the current study, text-internal diversity measures can provide a useful metric to contrast to other text-external sophistication measures. However, diversity measures only partially account for the differences across texts at different proficiency levels. To better explain these differences, other aspects of lexical breadth must be considered. As Cobb and Horst (2015, p. 194) write, "in our view, the extent to which L2 learner speech or writing contains diverse words regardless of their frequency (as in TTR-related measures) seems less revealing than the extent to

which it contains actual infrequent words." We now consider measures which focus specifically on infrequent words.

2.2.2 Lexical sophistication

Classically, lexical sophistication has been defined as the proportion of relatively advanced words produced by a learner in a text (Read, 2000), though more recently the construct has been operationalized in a number of ways (Crossley et al., 2015; Kim et al., 2018; Kyle, 2020; Kyle et al., 2018). To illustrate lexical sophistication, Meara and Bell (2001, p. 6) compared two sentences with the same number of tokens (5) and types (4):

1. *The man saw the woman.*
2. *The bishop observed the actress*.

Intuitively, these two sentences differ greatly in their level of sophistication, despite being equally diverse, due to the use of more advanced words like 'bishop' and 'actress'. Exactly what constitutes an advanced word, however, is a matter of some debate, though in general they are thought to be low-frequency lexical items (Laufer & Nation, 1995).

Regardless of how advanced words are defined, researchers in a variety of fields agree on the importance of lexical sophistication (Kyle et al., 2018): on average, higher proficiency learners produce texts with higher lexical sophistication (Crossley et al., 2019; Kyle & Crossley, 2017; Vögelin et al., 2019). Or seen from the opposite perspective, writers who use less frequent words are judged to be more proficient (Crossley & McNamara, 2012; Crossley & Skalicky, 2019). In studies that have measured both diversity and sophistication, sophistication better accounts for proficiency differences, especially at higher levels (Daller et al., 2003; Juffs, 2019). Measurement

and interpretation of sophistication remains a challenge (Daller et al., 2013), however, resulting in the propagation of related methods and tools including CLAN (MacWhinney, 2000), Coh-Metrix (Graesser et al., 2004), Lexical Complexity Analyzer (Lu, 2012), P_Lex (Meara & Bell, 2001), Lextutor (Cobb, n.d.), TAALES (Kyle et al., 2018), and PELITK (Naismith et al., 2022).

To measure and report lexical sophistication, the Lexical Frequency Profile (LFP; Laufer & Nation, 1995) and its web-based offshoot VocabProfile (Cobb, n.d.) have been the most widely used method over the last 25 years (e.g., Daller et al., 2013; Horst & Collins, 2006; Lindqvist et al., 2013; Morris & Cobb, 2004). After inputting a text, LFP outputs the percentage of words in each frequency band in an external corpus (e.g., the British National Corpus [BNC]) or wordlist (e.g., the Academic Word List [AWL; Coxhead, , 2000]). This reliance on data from external corpora is the current common practice in corpus linguistics, based on the premise that it is necessary to consult large corpora for 'empirical anchoring' in order to study smaller corpora (van Hout & Vermeer, 2007, p. 137).

Importantly, LFP considers only individual lexical items and is independent of syntactic or discoursal features of the text. Daller et al. (2013) provides a relevant example of LFP used to determine lexical sophistication. In this longitudinal study, the authors analyzed the vocabulary in essays from 42 students over the course of two years (294 essays total). One of the measures included was the number and percentage of advanced types, calculated using LFP. These data were used to report lexical sophistication as part of a larger effort to describe the learners' overall lexical proficiency. From this information, the authors were able to identify and model a latent learning curve of the learners' lexical development. A limitation of LFP is that it is not intended for use with texts shorter than 200 words. For this reason, another similar (though less frequently used) vocabulary profiler was developed, P_Lex (Meara & Bell, 2001), which employs the same

approach as LFP but with 10-word samples to produce a single sophistication metric. One major drawback of both LFP and P_Lex is that they do not take into consideration multi-word units (Lindqvist et al., 2013).

In the same Daller et al. (2013) study described above, a second measure of lexical sophistication was also calculated: the Advanced Guiraud (AG; Daller et al., 2003). Like LFP, AG requires frequency band counts as part of the formula $AG = (Advanced\ Types)/\sqrt{Tokens}$. AG also shares characteristics with the diversity measure TTR in that it is essentially a type-to-token ratio with the square root of the total tokens to reduce the sensitivity to text length. What makes AG a sophistication measure, rather than a diversity measure, is that only 'advanced types' are considered. These advanced types are identified by excluding the most frequently occurring 2000 lemmas (a word and its inflected forms), such as 'the' and 'be'. These lists are derived from corpora, e.g., the New General Service List (NGSL; Browne et al., 2013).

Importantly, not all frequency lists contain the same items because they draw from different corpora with different compositions (Nation, 2007). Consequently, AG scores will vary depending on the list used as the basis for advanced words (Naismith et al., 2018; Naismith et al., forthcoming; Tidball & Treffers-Daller, 2008). In Tidball and Treffers-Daller (2008), the authors compared AG scores of L2 French learners' oral production using three lists: a general frequency-based list, an oral frequency-based list, and a list based on teachers' judgments of basic items. Of the three, the teacher judgement list performed the best at differentiating between the proficiency groups, also outperforming the LFP in this regard. To contrast expert-speaker and learner corpus lists, Naismith et al. (2018) calculated AG for two learner populations using the basic words from the Corpus of Contemporary American English (COCA; Davies, 2008-) and the University of Pittsburgh English Language Institute Corpus (PELIC; Juffs et al., 2020; Naismith et al., 2022). The findings showed

that the frequency list from PELIC was better able to reveal differences between proficiency levels, more clearly depicting the lexical development of learners as they acquired new lexis in their studies. In a follow-up study, Naismith et al. (forthcoming) tested whether more local or global learner-corpus frequency lists would significantly impact AG scores across proficiency levels. In this case, it was determined that there was no appreciable difference in AG ranges, indicating that AG scores using learner corpus frequency data can be generalized to different contexts, regardless of the provenance of the learner corpus list. Overall, numerous studies incorporating AG have found it to be a reliable method for distinguishing between proficiency levels (Daller & Xue, 2007; Juffs, 2019; Milton, 2009; Tidball & Treffers-Daller, 2008).

In addition to LFP and AG, a variety of other sophistication measurements have been used in lexical studies. Notably, Crossley and his colleagues have operationalized sophistication by considering factors other than frequency, including a wide range of psycholinguistic metrics which relate to meaning, e.g., word concreteness, word imageability, and word familiarity (Crossley et al., 2012; Crossley et al., 2013; Crossley et al., 2011; Crossley & Skalicky, 2019; Guo et al., 2013; Kim et al., 2018; Kyle & Crossley, 2015, 2017; Kyle et al., 2018). This research direction is an exciting one, building upon previous findings in psycholinguistics relating to these same characteristics of words (see e.g., de Groot & Poot, 1997; Kroll & Merves, 1986; Tokowicz & Kroll, 2007). Nevertheless, it is important to avoid discrediting the previous form-based metrics: the new metrics are not better, but rather measure different aspects of sophistication (Lenko-Szymanska, 2019). As well, how advanced words are categorized may be partly discipline-specific, and there is a strong overlap between the many lexical sophistication indices (Kim et al., 2018). Non-frequency-based metrics will not be considered further here as they are not being used in the study. Instead, the pragmatic course of action suggested by Cobb and Horst (2015) will be

followed: using a blend of text-internal diversity measures (vocD) and text-external sophistication measures (AG).

2.2.3 Low-, mid- and high-frequency

Word frequency is an integer variable, but for practical reasons it is commonly partitioned into frequency bands. For example, with the original LFP, words were sorted, in part,[5] according to whether they were in the 1000 most frequent words (K1) or next 1000 most frequent words (K2) in the General Service List (GSL; West, 1953). Using the LFP categories, Biber and Gray (2013a) found that in their sample of speech and texts from TOEFL test takers, 85% of words were from K1 and that higher-level responses used more K2 and AWL words.

At an even broader level, frequency bands have been clustered under the labels *low-frequency* and *high-frequency*. This simple classification is thought to enable stakeholders to make quick cost/benefit analyses as to whether a word is deserving of classroom attention, with high-frequency words potentially more useful in a variety of contexts compared to low-frequency words (Nation, 2011). As a result, high-frequency words can be said to have greater text coverage, i.e., they account for a greater percentage of all the words found in a text (Cobb & Laufer, 2021). Coverage, therefore, is typically the quantitative metric used to determine the border where high-frequency words end and low-frequency words begin. With each additional K-band (1000 words), coverage becomes increasingly small (Laufer & Ravenhorst-Kalovski, 2010; Schmitt & Schmitt, 2014). Thus, 'high-frequency' has traditionally been defined as K1-2 (Schmitt & Schmitt, 2014;

[5] LFP also categorized the percentage of words in the AWL and the percentage of words not in any of the other defined categories (off-list). Consideration of academic and technical vocabulary lists are not considered here.

Vilkaitė-Lozdienė & Schmitt, 2020), providing coverage of approximately 80% of texts (Nation & Waring, 1997).[6] The use of the high/low frequency dichotomy has also yielded results when considering learner production. For example, in Lindqvist et al. (2011) compositions from the more advanced learner group contained a significantly higher proportion of low-frequency words.

Many other authors have suggested a more nuanced three-way distinction between low-, mid-, and high-frequency lexical items (Dang, 2020; Naismith & Juffs, 2021; Nation, 2016; Schmitt, 2010; Vilkaitė-Lozdienė & Schmitt, 2020). In this format, the K1-2 frequency bands are high-frequency, K3-9 are mid-frequency, and K10+ are low-frequency. Descriptively, mid-frequency lexis can be thought of as wide-ranging, moderately frequent general-purpose vocabulary (Nation & Anthony, 2013). From a coverage perspective, the mid-frequency categorization is a sensible one. In order for learners to comprehend a text, they must typically know 95-98% of the words in it (Hu & Nation, 2000; Nation, 2006; Schmitt et al., 2011), i.e., a 'language threshold' of one unknown word per 50 (Hu & Nation, 2000). Clearly, this threshold is far greater than the 80% coverage provided by high-frequency words. For comprehension of authentic texts needed for university study, the 95-98% threshold equates to knowledge of approximately the K8-9 frequency bands (Laufer, 1989; Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006). After this point, there are minimal coverage differences with each subsequent K-band, e.g., +0.32% from K10 to K11 or 0.10% from K14-15 in COCA (Schmitt & Schmitt, 2014).

In creating a mid-frequency category, lexical items can be identified which are pedagogically useful (Nation & Anthony, 2013; Schmitt, 2010), especially for learners wishing to

[6] Schmitt and Schmitt (2014) argue that the 'high-frequency' threshold should be K3 based on the coverage needed to 'largely understand' conversational English, but this practice has not yet been widely adopted.

study in an L2 academic environment (Vilkaitė-Lozdienė & Schmitt, 2020). These lexical items are distinct from, but overlap with, academic vocabulary lists such as the AWL (Coxhead, 2000) or AVL (Academic Vocabulary List; Gardner & Davies, 2014). This more precise learning goal (Dang, 2020) is a logical next step for learners who know the high-frequency words (Nation, 2016), and it can lead to clear rewards for learners in terms of academic success and enjoyment of authentic texts (Schmitt & Schmitt, 2014). Critically, to date, there is no research investigating the effects of productive use of mid-frequency lexis compared to low-frequency lexis in terms of readers' perceptions of text quality. The claims by Vilkaitė-Lozdienė and Schmitt (2020) of academic usefulness therefore need to be validated in terms of productive usage to determine whether there is any additional benefit to learning low-frequency words.

2.2.4 Counting units

Up to this point, frequency has been discussed without much attention paid to what exactly is being counted, with mention of words, word families, and lemmas. However, it is important not to gloss over this methodological concern as the choice of counting unit greatly affects the word-selection process, and by extension, corpus frequency lists and measures of lexical richness (Brezina & Gablasova, 2015; Kyle, 2020). The most simplistic form of counting would of course be to consider unique surface word forms, i.e., *types*. However, doing so ignores the fact that the singular form of a noun, like 'turtle', and the plural form, 'turtles', are intimately related. Intuitively, we would not consider a learner who knows 'turtle' and 'turtles' to have the same lexical breadth of knowledge as a learner who knows 'turtle' and 'tortoise'. We must therefore consider other potential units which group types together in such a way that better represents vocabulary knowledge.

One such grouping unit is the *word family*, advocated for by Bauer and Nation (1993). For example, the 'happy' word family would include 'happy', 'happiness', 'unhappy', and 'happier', i.e., a number of forms related to 'happy' either through inflection or derivation. Bauer and Nation (1993) justify this grouping on the assumption that it reduces learning burden, so that once one form is known, the others will not have to be learned separately. A word family is therefore "a base word and all its derived and inflected forms that can be understood by a learner without having to learn each form separately" (Bauer & Nation, 1993, p. 253). Critically, in this definition, it is necessary to define and determine what "can be understood", which traditionally required curation of word family lists using criteria such as frequency, regularity, and productivity. More recently, Cobb and Laufer (2021) have sought to standardize the methodology defining what constitutes a word family, leading to their own Nuclear Word Family List (NFL7). In this list, word families are refined to include only forms which account for 7% or more of the total tokens of the family, thereby balancing considerations of size and coverage. It remains to be seen whether this new approach to word families will be widely adopted in the research community.

Even using carefully selected word families, a number of problems arise. For example, Lextutor uses family groupings by Paul Nation from the combined 25k BNC-COCA lists[7]. Here, under the headword 'act', we find 'actresses', 'actionable', and 'inaction'. Certainly, for some learners the derivational links between the forms will allow for comprehension, but it is unclear to what extent this is case. As Brezina and Gablasova (2015) describe, there is an assumption of transparency between these forms of varying semantic distance which is problematic, especially considering the variability of learners' morphological awareness. Friedline (2011), for example, found that students had difficulty identifying words related by derivation, regardless of their

[7] https://www.lextutor.ca/vp/comp/coca.html

proficiency level and L1. Productively, Schmitt and Zimmerman (2002) observed that learners showed an overreliance on one or two derivational forms. It is therefore debatable the degree to which knowledge of one word family member leads to knowledge of others sight unseen (Gyllstad, 2013; Vilkaitė-Lozdienė & Schmitt, 2020).

A widespread and more conservative approach to grouping words is to use *lemmas* as the counting unit (Brezina & Gablasova, 2015; Gyllstad, 2013; Lindqvist et al., 2011; Lindqvist et al., 2013; Schmitt, 2010). A lemma consists of a word and its inflected forms (but not other derivations), such that the verb forms 'act', 'acted', and 'acting' would be one unit, but 'act'' and 'actionable' would be distinct. Dictionaries are typically organized by lemma (i.e., the headwords), and lemma information is helpfully provided with many lexical resources, including the Brown Corpus, the New-GSL (Brezina & Gablasova, 2015), and COCA (Davies, 2008-). It is certainly easier to compute the lemmas for a given text than the word families, and it may be considered less arbitrary as well (Vilkaitė-Lozdienė & Schmitt, 2020). However, to a lesser extent, the issue of transparency between forms may still occur with lemmas, e.g., for irregular plural forms like 'child' and 'children'.

Ultimately, the choice of whether to use lemmas or word families as the counting unit might best be decided based on the type of data and research questions being investigated (Webb, 2021). For studies of receptive use, word families could give a better sense of the number of units needed to achieve certain levels of coverage or comprehension as a reader may infer meanings of unknown derivations. In contrast, for describing production data, as in this study, lemmas are a more apt counting unit (Dang, 2020; Nation, 2007; Vermeer, 2004). For one, lemmas within the same word family can associate with different collocations and grammatical constructions. What

is more, using lemmas avoids the potential issue of crediting learners with knowledge of family members of which they have no productive knowledge.

As it pertains to the cut-off points for the low-, mid-, and high-frequency categories, the choice of counting unit has surprisingly little effect. In their empirical exploration of this issue, Vilkaitė-Lozdienė and Schmitt (2020) describe how for the K1-3 bands, using lemmas still results in 98% text coverage, signifying the same threshold for word families and lemmas. For the mid-frequency/low-frequency border, the findings again point to a relatively small difference in coverage regardless of unit. Schmitt and Schmitt (2014) posited that K10+ word families be considered low-frequency. Based on coverage, this figure would translate to K11+ for lemmas (Vilkaitė-Lozdienė & Schmitt, 2020), i.e., slightly more. However, establishing the boundary for 'low-frequency' is notoriously difficult given the small increases in coverage and differences in frequencies across corpora at these ranges. As well, it is estimated that for academic texts specifically, the number of lemmas required for comprehension is somewhat less than for other authentic genres like adult fiction or magazines.

In sum, when calculating vocabulary knowledge of text sophistication, the choices of counting units and thresholds are important ones for producing reliable results. In this study lemmas are used as the counting unit based on the type of data (productive), the aspect of lexis being researched (collocations), and the format of the external datasets being consulted. In keeping with common practice based on coverage statistics, the category thresholds are K1-2 for high-frequency, K3-9 for mid-frequency, and K10+ for low-frequency.

2.3 Lexical accuracy

A third aspect of lexical proficiency that is relevant to the proposed study is *lexical accuracy*. Lexical accuracy can be considered part of lexical depth, and simply put, it is the ability to produce writing free from lexical errors. The evidence for the importance of lexical accuracy is overwhelming. In general, there is a strong negative correlation between number of errors and holistic ratings (Polio & Shea, 2014), and in terms of quantity, lexical errors have been found to occur more than grammatical errors (Agustín Llach, 2007, 2011; James, 1998; Qian & Lin, 2020). What is more, in terms of *error gravity* – the likelihood of an error being disruptive to communication (Rifkin & Roberts, 1995) – lexical errors typically have a greater impact (Agustín Llach, 2007; Ellis et al., 1994; Khalil, 1985; Santos, 1988), though see Agustín Llach (2007) for an exception. When studies developing models of lexical proficiency do not consider accuracy as a variable, there is often acknowledgment of this limitation (e.g., Crossley & McNamara, 2012; Kyle & Crossley, 2015).

Because lexical errors affect communication, they are highly prominent (Fritz & Ruegg, 2013; Hawkey & Barker, 2004) and are therefore judged more severely by readers and listeners (Ellis, 2008; Santos, 1988). As a result, lexical errors affect academic achievement (Daller et al., 2003; Engber, 1995; Hawkey & Barker, 2004) and pose a great challenge to L2 writers (Xie, 2019). Lexical errors can also be used by researchers, in conjunction with other lexical metrics, to measure the quality of written work (Agustín Llach, 2007, 2011; Council of Europe, 2001; Ginther & Grant, 1997; Hawkey & Barker, 2004; Lee, 2003; Polio, 1997). For example in Ginther and Grant (1997), the authors compared error counts in student texts at two proficiency levels. Whereas as only 1.5% of the higher-level group's errors were word-choice errors, this type of lexical error accounted for 4.4% of the lower-level group's errors. And yet, research into lexical accuracy is minimal in

comparison to the other aspects of lexical proficiency already described (Agustín Llach, 2007; Granger, 2003), despite the importance of helping learners to improve their accuracy. In writing, accuracy takes on even greater significance than in speech because written errors are more likely to be an indication of a deficit in lexical competence – in speech, more errors may be attributed to a problem with performance due to the demands of speech production in real time (Lenko-Szymanska, 2019; Perfetti & Hart, 2002).

Although the importance of lexical accuracy is apparent, determining what exactly constitutes an error is not a straightforward matter (Ellis, 2008; Engber, 1995; Polio & Shea, 2014). Here I adopt the broad definition of Agustín Llach (2011, p. 75), which she proposed after a thorough review of studies in this field: "A 'lexical error' is a deviation in form and/or meaning of a target-language lexical word." This definition allows for inclusion of different types of errors which might impede communication, ranging from spelling mistakes (form) to word choice (meaning). Even with a clear definition, identifying errors is no simple matter (James, 1998), and classifying/annotating them is a time-consuming process requiring a number of judgment calls. For example, Ruegg et al. (2011) found it a challenge to determine what exactly constitutes a lexical error vs. a grammatical error, and Lenko-Szymanska (2019) questioned whether or not pragmatic and spelling errors should be included as lexical errors. Ultimately, different studies have categorized lexical errors using taxonomies of their own creation which suit their particular research questions, in line with James' (1998) recommendation that error description systems be as well-developed as possible, yet still self-explanatory and user-friendly.

The simplest categorization is to simply distinguish between grammatical and lexical errors (Ellis, 2008). Adopting a slightly more fine-grained approach, Ruegg et al. (2011) considered lexical errors relating to 'idiom usage', 'contextually appropriate word choice', 'word class', and

'spelling which impeded meaning'. In a much more detailed error taxonomy, Granger (2003) devised a system with 9 error domains and 36 categories, so that, e.g., in the 'Lexis' domain there are categories for 'meaning', 'verb complementation', etc. In a review of this field, Agustín Llach (2011) noted that lexical error categories commonly used by researchers include word choice, omission, unusual word forms, word order, borrowings, lexical creations, and spelling.

Unsurprisingly, given this range of lexical error taxonomies, the findings are similarly varied though many common error types can be subsumed under the umbrella of 'word choice'. For example, both Ginther and Grant (1997) and Hawkey and Barker (2004) found that word choice errors were the most impactful, which corresponds to the notion that such errors have high gravity. Similarly, in her study of L2 Spanish young learners, Agustín Llach (2007) found that the most impactful error types (negatively correlating to text ratings) were borrowings and coinages, which can be considered types of word choice errors stemming from the learners' L1s. Interestingly, in this study, misspellings were the most common error type (74.8%) though not the most impactful, whereas in an earlier study, Grobe (1981) did find spelling errors to be one of the two best predictors of text ratings (along with lexical diversity). Finally, in a recent meta-analysis of error analysis studies, Xie (2019) reviewed 34 papers focusing on high-frequency or high-gravity errors. From these, a list of the 31 most important linguistic error types was compiled. This list includes seven critical lexical error types with collocation errors ranked #1 overall (another type of word choice). In addition, prepositional errors were classified as syntactic errors and ranked #1 amongst the syntactic error types, but these can arguably be considered lexical word choice errors as well.

While error taxonomies vary greatly across studies, error metrics are more consistent. The typical quantitative approach is to count either error free units, like T-units or clauses (Ellis &

Yuan, 2004; Polio, 1997), or errors themselves (Agustín Llach, 2011; Engber, 1995; Ginther & Grant, 1997; Linnarud, 1986). These counts can then be normalized as a ratio, e.g., number of errors per word, per lexical word, or per 100 words. (See Wolfe-Quintero et al. (1998) for a thorough overview of accuracy measures across studies.) Counting errors (or absence of errors) allows, theoretically, for more objective measurement of accuracy, though it does not provide information about error gravity (Polio, 1997). In contrast, using holistic scales to assess accuracy allows for consideration of quality as well as quantity, though at the cost of inter-rater reliability (Polio & Shea, 2014; Polio, 1997). In an analysis of the reliability of these types of accuracy measures, Polio and Shea (2014) did not uncover any evidence that one accuracy measure was more valid than the others.

2.4 Chapter summary

In summary, the above review of error taxonomies and metrics presents two implications for studies investigating lexical accuracy: (1) it is imperative to clearly define what is considered an error based on the focus of the study; and (2) having identified and categorized the errors, frequency counts or ratios can be used to describe the lexical accuracy of the texts. More broadly in this chapter, the main takeaway is that learners' lexical proficiency can be observed, in part, through the lexical sophistication and accuracy of their texts. However, it is necessary to carefully consider how these dimensions of lexical proficiency are operationalized and measured. This study targets both sophistication and accuracy in estimating which factors influence examiner ratings. Having established the basics of lexical proficiency, the next chapter considers one component of this construct: collocational proficiency.

3.0 Collocational proficiency

This chapter starts with a brief overview of *formulaic sequences* before turning to one specific type, collocations. It then considers different ways of conceptualizing, identifying, and measuring collocation use in writing. Finally, we review studies that have looked at L2 collocation use and how this differs from L1 collocation use.

3.1 Formulaic sequences

Sinclair et al. (2004, p. 81) wrote that "the lexical unit is best described maximally, not minimally", building on Firth's (1957, p. 11) famous proclamation that "you shall know a word by the company it keeps!" That is to say, to truly understand language use, we must consider lexical units above the word level. What exactly these units are and what to call them has long been discussed, with popular terms including *multiword expression/sequence/item/unit* (see Wray, 2002 for a full discussion). Here, I use the umbrella term *formulaic sequence* (FS) which best aligns with this study's research methodology: FS refers to any string "perceived by the agent (i.e. learner, researcher, etc.) to have an identity or usefulness as a single lexical unit" (Siyanova-Chanturia & Pellicer-Sánchez, 2020, p. 6). FSs can therefore include phrasal verbs (e.g., 'bump into'), idioms (e.g., 'spill the beans'), lexical phrases (e.g., 'not only X but Y'), and collocations (e.g., 'single parent') (Wood, 2020).

This viewpoint sees FSs as being at least somewhat holistic in nature, consisting of a single lexical chunk (Nattinger & DeCarrico, 1992; Sinclair, 1991; Uchihara et al., 2021; Wood, 2002;

Wray, 2002) with psycholinguistic validity (Jiang & Nekrasova, 2007; Schmitt et al., 2004), though the exact extent to which these entrenched chunks are holistically stored and retrieved continues to be investigated (Siyanova-Chanturia, 2015). As such, the earlier discussion of lexical diversity, sophistication, and accuracy applies equally to FSs.

The most common description of FSs is that they are 'pervasive' in language (e.g., Bestgen & Granger, 2014; Conklin & Schmitt, 2012; Nattinger & DeCarrico, 1992; Sinclair, 1991; Siyanova-Chanturia & Martinez, 2015; Siyanova-Chanturia & Pellicer-Sánchez, 2020). As noted, they can take many forms and may vary in length, figurativeness, compositionality, and other characteristics (Elgort & Siyanova-Chanturia, 2021; Lewis, 1993; Siyanova-Chanturia & Omidian, 2020; Siyanova-Chanturia & Spina, 2020). What makes FSs so pervasive is their frequency (Siyanova-Chanturia & Pellicer-Sánchez, 2020). Across studies, FSs have been estimated to account for 20 to 50% of all expert speaker speech and writing (Conklin & Schmitt, 2012; Hill, 2000; Siyanova-Chanturia & Martinez, 2015). Therefore, FSs can be considered 'basic linguistics units' (Durrant, 2019, p. 211), in-line with theoretical approaches including the idiom principle (Sinclair, 1991), pattern grammar (Hunston & Francis, 2000), and construction grammar (Goldberg, 2006). Knowledge of FSs has been shown to be a critical component of communicative competence (Henriksen, 2013), fluency (Nation, 2013; Uchihara et al., 2021), and language processing (Nattinger & DeCarrico, 1992; Siyanova-Chanturia, 2015; Wolter & Yamashita, 2018). Due to their importance, there has been extensive research into FSs in psycholinguistics and applied linguistics, particularly since the 1990s (Henriksen, 2013; Öksüz et al., 2021; Siyanova-Chanturia & Omidian, 2020; Siyanova-Chanturia & Pellicer-Sánchez, 2020; Vilkaitė, 2016).

3.2 Collocations

Of the many types of FSs, collocations are the most commonly investigated in lexical research (Gyllstad & Schmitt, 2019; Henriksen, 2013; Vilkaitė, 2016). Since the 1930s (Palmer, 1933) the need for collocational knowledge has been recognized, and there is now broad consensus of its important role in overall proficiency. Popular frameworks of lexical knowledge typically depict collocations as a unique dimension (e.g., Dóczi & Kormos, 2016; Nation, 2013; Read, 2004). Relevant to this study, collocations are especially frequent in academic discourse (Ellis & Simpson-Vlach, 2009; Hyland, 2012).

3.2.1 Collocation identification

Pinpointing what exactly constitutes a collocation is a challenging endeavor (Read, 2000). In language classrooms, teachers often define collocations along the lines of "words that are found together in language" (British Council, n.d.), echoing Sinclair's description of collocations as "typically regular predictable combinations" (2004, p. 21). These combinations may be immediately adjacent as in 'illegally parked' or 'speak English' (Bestgen & Granger, 2014) or separated, often by functional words, as in 'bread and butter' or 'drink a beverage' (Church & Hanks, 1990). However, in lexical research what is designated as a collocation differs to varying degrees across studies. For instance, Cowie and Howarth (1996) refer to 'semantically opaque units', whereas Laufer and Waldman (2011) list 'relative transparency of meaning' as a criterion. In general, however, there are two common approaches for identifying collocations: a phraseological approach and a frequency-based approach, each with inherent strengths and limitations (Lundell & Lindqvist, 2012; Siyanova-Chanturia & Omidian, 2020).

In the phraseological approach, collocations are identified based on syntactic, semantic and pragmatic linguistic criteria (Henriksen, 2013; Lundell & Lindqvist, 2012). For example, researchers in this tradition often differentiate between *free combinations* (which are not collocations) and *restricted collocations*; free combinations are compositional in nature and there are numerous plausible replacements for the words (Howarth, 1998; Lee, 2019; Nesselhauf, 2005). Wolter (2020) gives the example of 'pay the bill' versus 'pay attention'. Here, 'pay the bill' is a free combination due to the literal meaning and the option to replace 'bill' with a large number of other unrelated nouns. In contrast, 'pay attention' is a restricted collocation because the verb is more figurative and the subsequent possible nouns are more restricted and semantically related. Exemplifying this approach, Nesselhauf (2005) extracted verb-noun collocations from a learner corpus using multiple methods. First, all potential combinations were checked against four dictionaries. They were also checked against the British National Corpus (BNC) to see if that same identical form-meaning pattern was present at least five times. Finally, for any remaining items, native-speaker acceptability judgements were consulted. In favor of this approach, word combinations can be extracted which have clear semantic relations between the words, even if those word combinations appear with only low frequency in a reference corpus (Henriksen, 2013; Howarth, 1998). There are dangers as well to relying on human intuition to identify FSs, as judgments may be inconsistent across judges. After all, individuals' experience of language varies so that perceptions of frequency and salience are not uniform. Judges may also be affected by factors such as fatigue (Wray, 2002).

In a frequency-based view of collocations, the probability of co-occurrence of words is of paramount importance (Henriksen, 2013), but semantics are not usually a factor (Macis & Schmitt, 2016). For example, in an early definition, Nattinger and DeCarrico (1992) describe collocations

as a node word which occurs in a given span at a greater-than-chance frequency. Definitions such as these are undeniably objective and therefore practical for replication and comparison of findings. However, there is no consideration of factors such as memory storage or processing, and word combinations may reach collocation status which do not in fact have psycholinguistic validity (Henriksen, 2013).

A combination of phraseological and frequency-based approaches is also possible as there is overlap between these conceptions of collocation (Evert, 2009). Many researchers use this combined approach, for example by starting with computational extraction and then subsequently applying phraseological criteria (Henriksen, 2013; Laufer & Waldman, 2011; Macis & Schmitt, 2016; Naismith & Juffs, 2021). In adopting a combined approach, it is possible to take into account 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):

> [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, combinations which are 'habitually occurring' can be measured statistically using a frequency-based approach and 'restricted co-occurrence and relative transparency of meaning' can be determined from a phraseological perspective. As part of both frequency-based and combined approaches, it is therefore necessary to carefully consider frequency metrics and the distance between the collocating words. We will now consider each of these in turn.

3.2.2 Collocation association measures

Pure frequency counts can be used to determine common word combinations, but these tallies do not capture the strength of association between the words. For instance, sequences like 'I do not' and 'there is a' are highly frequent in a learner corpus (Vercellotti et al., 2021), but there are clearly no strong lexical bonds between the words. As a result, corpus linguists have devised association measures based on the concept that words in collocations are 'mutually expectant' (Firth, 1957). Specifically, measures have gained popularity in the collocation research community that are believed to correlate to human intuitions. Amongst these, two popular association measures are Mutual Information (MI) and t-score, which describe collocational strength based on co-occurrence between words, though of the two MI has become the "field-standard measure" (Öksüz et al., 2021, p. 61; cf. Kang, 2018 for a critique). For both metrics, a higher score indicates that the words are more likely to co-occur compared to chance. A typical convention based on previous research is to consider words with an MI score over 3 or a t-score over 2 to be a collocation (Church & Hanks, 1990; Hunston, 2002; Jiang, 2009). Other types of measures of formulaicity not described here include psycholinguistic measures (e.g., reaction times) and acoustic measures (e.g., phonological coherence) (Wood, 2020).

Considering MI, in Simpson-Vlach and Ellis (2010), the authors elicited teachers' ratings of FSs in terms of their formulaicity and pedagogical value. These ratings were then compared to statistical measures including MI and raw frequency. The study found that MI better predicted teachers' judgements than raw frequency, indicating that MI corresponds more closely to human intuitions of formulaicity and 'teaching worth'. Unlike other metrics, MI prioritizes less common words that are typically found together, e.g., 'furrowed brows', especially when based on data from large corpora (Evert, 2009). For this reason, when using MI it is recommended to include a

minimum frequency threshold (Davies, 2008-; Evert, 2009), commonly from 5 to 10 (Biber & Gray, 2013a; Granger & Bestgen, 2014; Simpson-Vlach & Ellis, 2010; Wood & Namba, 2013). MI may also be unreliable for multiword associations and is normally only reported for two-word combinations (Hyland, 2012). In contrast to MI, t-scores give greater weight to frequency, as with collocations composed of high-frequency words which may also be found in a wider range of contexts, e.g., 'good work' (Bestgen & Granger, 2014; Durrant, 2019; Granger & Bestgen, 2014).

One other important methodological consideration when determining collocations is *collocation span* (Evert, 2009). Collocation span, or 'window', refers to the number of words on each side of a node word. In (1), for the node word 'skills', the underlined span is two. Thus, when considering potential collocations, 'taught… skills' would be considered, but not 'skills… adult'.

1. These children are taught necessary **skills** for survival as an adult from a very early age.

Collocation spans matter because the greater the distance from the node, the fewer significant collocates there are (Sinclair et al., 2004). The most common span ranges have been listed as 4 (Durrant, 2014), 1-4 (Sinclair et al., 2004) and 3-5 (Evert, 2009), e.g., studies with spans of one (Sinclair, 1991), two (Naismith & Juffs, 2021), three (Biber & Gray, 2013a), four (Mollin, 2009), and five (Jiang, 2009). Using a span of 1-4 aligns with the processing advantage found by (Vilkaitė, 2016) for collocates separated by up to three words.

3.2.3 Between grammar and lexis

Collocations occupy an interesting intermediary space on many linguistic clines, somewhere between literal and figurative, compositional and holistic, and grammatical and lexical. As Nattinger and DeCarrico (1992, p. 1) write, chunks of language like collocations can be

considered lexico-grammatical units which "exist somewhere between the traditional poles of lexicon and syntax". For example, the collocation 'tell (me) a story' contains a clear lexical component, with 'tell' and 'story' strongly associated. There is also the grammatical structure of ditransitive verb + (indirect object) + direct object.

Usage-based approaches to language are particularly well-equipped for understanding these types of linguistic structures (see Section 2.2 for a description of usage-based principles). Collocations, which are a frequent type of FS, are more likely to be encountered, thereby increasing the likelihood that they are noticed and stored in memory. In fact, it is possible that FSs can be acquired starting with individual words which are then combined to form longer sequences (Ellis, 1996), and it is also possible for words or syntactic patterns to be extracted from known FSs (Tomasello, 2003). In other words, although FSs may originally be non-compositional chunks, as a learner increases in proficiency, they may analyze these chunks and learn to use elements from these constructions independently.

This concept of *constructions* is central to the usage-based framework of Construction Grammar. In essence, constructions are form-function pairings (Goldberg, 2006; Goldberg, 2013) and range from more concrete constructions like words or idioms, to more abstract constructions like the passive voice (Goldberg, 2013). From an early age, learners develop an inventory of such constructions (Tomasello, 2000), which can be strengthened through repeated exposure and use and can also be generalized to novel contexts as they become less lexical and more syntactic in nature (Bybee, 2006, 2010). In a series of experiments, Gries and Wulff (2005) found evidence to support the notion that for L2 English learners constructions have a distinct ontological status.

From this perspective, the distinction between grammar and lexis is therefore an artificial one because the two are interdependent (Römer, 2009; Siyanova-Chanturia & Martinez, 2015) and

on a spectrum of construction types (Halliday & Matthiessen, 2014). Collocations fall somewhere in the middle of this spectrum as they possess both concrete (vocabulary-like) qualities and abstract (grammar-like) qualities. In line with this perspective, Milton (2013, p. 75) observed that learners' vocabulary development "[meshed] very closely" with grammatical development, and Engber (1995) noted that grammatical and lexical errors overlap. Discussing assessment implications, Ruegg et al. (2011) suggest that a single lexicogrammar category is more valid for rating writing based on the difficulty in separating lexis and grammar, a suggestion supported by Römer (2017) with respect to speaking assessment.

More commonly, however, grammar and lexis are still considered to be discrete categories, even if they are "inextricably intertwined" (Qian & Lin, 2020, p. 66). This distinction remains useful for assessment purposes (Weigle, 2002) and has face validity since teachers typically adhere to this division in their classroom teaching (Parr & Timperley, 2010), and it is the norm in coursebooks. Some researchers have also proposed, based on production data, that lexical and grammatical complexity are unique dimensions of L2 performance and proficiency (Bulté & Housen, 2014; Foster & Tavakoli, 2009; Skehan, 2009). For example, Foster and Tavakoli (2009) suggest that intermediate learners rely more on lexical processing before integrating lexical and grammatical processing at higher levels of proficiency, As such, analytic scales must continue to contend with clarifying how elements of language are to be classified and assessed.

3.3 L2 collocation use

Research has consistently shown that expert speakers and learners differ in their knowledge and use of collocations and that collocations can be used to distinguish between these two

populations (Granger & Bestgen, 2014). These differences are most apparent in learners' production as opposed to comprehension (Siyanova-Chanturia & Sidtis, 2019). On the one hand, learners overuse collocations that they prefer or know well, what Hasselgren (1994, p. 237) coined "lexical teddy bears" (De Cock et al., 1998; Fan, 2009; Granger, 1998; Hasselgren, 1994; Laufer & Waldman, 2011; Li & Schmitt, 2009). For example, in an investigation of adjective+adverb collocations, Granger (1998) observed that learners overly relied on the high-frequency adverb 'very' as opposed to other lower-frequency, less generalizable adverbs.

On the other hand, learners' underuse of collocations is well documented (Altenberg & Granger, 2001; Chen & Baker, 2010; Durrant & Schmitt, 2009; Fan, 2009; Granger, 1998, 2019; Howarth, 1998; Laufer & Waldman, 2011; Nesselhauf, 2005), and they use less variety of collocations as well (De Cock et al., 1998; Tsai, 2015). Specific types of underused collocations include adverb+adjective (Granger, 1998; Lorenz, 1999) and verb+noun combinations (Laufer & Waldman, 2011; Nesselhauf, 2005). There is also an L1 effect for collocation use, with L1 transfer shown in many studies for a variety of L2s (Fan, 2009; Granger, 1998; Jiang, 2009; Lee, 2019; Nesselhauf, 2005). Learners prefer collocations with L1 equivalents, i.e., 'congruent' collocations which are acceptable in the L1 and have been translated (Granger, 1998; Jiang, 2009; Lee, 2019).

Students, teachers, and researchers agree that collocations (and FSs more generally) are a challenge to acquire (Siyanova-Chanturia & Schmitt, 2008; Siyanova-Chanturia & Spina, 2020; Wray, 2002) and that knowledge of collocations trails that of single words (Granger, 2019; Laufer & Waldman, 2011). In fact, collocational mastery is often never reached, remaining an issue at even advanced levels (Henriksen & Stæhr, 2009; Lundell & Lindqvist, 2012), more so with production than comprehension (Linnarud, 1986). The reason for such failure is likely a combination of factors which may include collocations' relative infrequency in input (Gyllstad &

Wolter, 2016), the lack of a literal counterpart in the learner's L1 (Macis & Schmitt, 2016), or their lack of salience as linguistic items (Lee, 2019; Wolter, 2020). Poor collocation knowledge could also be the result of the manner in which they were taught (Wray, 2009), with many students learning vocabulary as single words with little attention to collocational properties (Jiang, 2009; Siyanova-Chanturia & Spina, 2020).

An effective pedagogical focus on collocations is therefore worthwhile given their impact on readers/listeners. Collocational knowledge is one factor which can distinguish between advanced learners and lower levels of proficiency (Ha, 2013; Lundell & Lindqvist, 2012). In an experimental study, Lee (2019) observed this proficiency effect as higher proficiency learners were more likely to reject unacceptable collocations than lower proficiency learners, but were still equally likely to accept acceptable collocations. In a corpus study, Kim et al. (2018) found that greater use of strongly associated bigrams and trigrams (i.e., collocations) corresponded to higher proficiency writing. It is therefore a useful measure of development (Lundell & Lindqvist, 2012), supporting findings that more use of formulaic language correlates with higher academic achievement (AlHassan & Wood, 2015; Jones & Haywood, 2004). In sum, the use of lexical analyses that go beyond the single word are necessary to truly capture the development of learners' lexical proficiency (Kim et al., 2018; Read & Nation, 2006).

Knowledge of collocations can also aid learners in developing other aspects of language proficiency; there is less learning burden/more fluent processing if items are stored as chunks, providing scaffolding for learners to understand and produce academic texts (Durrant, 2019). This hypothesis is supported by psycholinguistic studies such as Sonbul (2015) which found that learners' sensitivity to FSs increases with proficiency. Likewise, in the longitudinal Garner and Crossley (2018) study, the researchers tracked learners' bigram and trigram usage over four

months and saw significant growth, especially bigram growth for beginners. Taken together, these findings support the position of Wray (2002, 2018) who describes more broadly how FSs play a different role in L1 and L2 learning, processing, and use, with learners relying less on the use of FSs.

3.3.1 Collocational sophistication

To discuss collocational sophistication, it is necessary to discuss collocation frequency. However, there are two distinct types of frequency. One approach is to count whole collocations, what I term a *collocation frequency approach*. For example, in COCA the lemma combination of 'say' and 'needless' (as in 'needless to say') occurs 3,735 times (span = 4) and has an MI of 4.37. As such, it is the 251st most common lemma combination and can be considered an extremely high-frequency collocation. If we then look at the lemmas (or words) individually, what I call a *collocate frequency approach*, we see that 'say' is certainly high frequency (lemma frequency = 4,096,416, lemma rank = 26, band = K1). However, this is not the case for 'needless' (lemma frequency = 4,942, lemma rank = 8,468, band = K9), which would be considered mid-frequency, verging on low-frequency. If a learner uses 'needless to say' in an essay, should this collocation be taken as evidence of low sophistication because it is a common collocation or high sophistication because it is a collocation containing an uncommon word?

These two approaches to viewing collocational sophistication reflect the nature of collocations discussed in Section 3.1 – they are simultaneously holistic chunks and also compositional strings of words. Processing studies in this area support this view. In Wolter and Yamashita (2018), L1 and advanced L2 groups completed an online collocation acceptability judgment task. It was discovered that both groups' processing speeds were affected by word

frequency and collocational frequency simultaneously, indicating multiple levels of representation. The L2 group was also more affected by word frequency than the L1 group, signaling lesser collocational sensitivity. Öksüz et al. (2021) conducted a similar subsequent study but extended their investigation to look at the interaction of word frequency and collocational frequency effects. As in Wolter and Yamashita (2018), both L1 and advanced L2 groups were sensitive to word-level and collocational frequency, but unlike Wolter and Yamashita (2018), the two groups were equally sensitive to word frequency. Based on these findings, the authors concluded that "repeated use of multiword sequences leads to growing prominence of the whole sequence, but the information about the parts is still accessible" (Öksüz et al., 2021, p. 89).

Of the two, the collocation frequency approach is more common. One of the principal findings of studies in this vein is that the correlation between collocation frequency and collocation knowledge is not robust. In a meta-analysis of 19 collocation studies, Durrant (2014) found that frequency correlated only moderately with collocation knowledge; other important factors included semantic transparency and the amount of social engagement of learners. As there are far fewer occurrences of any given word combination than the component words, from a usage-based perspective it is not surprising that this relative lack of exemplars will not have as great an effect on the acquisition of collocations (Macis & Schmitt, 2016). We might therefore expect that for collocation learning, it is not just the number of occurrences in a corpus that matters, but also the kind of individual exposure that a learner receives.

Studies using a collocate frequency approach are more interested in collocations in relation to the individual collocates contained within. For example, Ebrahimi (2017) investigated the collocational knowledge of Iranian EAP learners, specifically the K1 band from the new-GSL list (Brezina & Gablasova, 2015). A major component of the battery of tests looked at collocational

knowledge of these lemmas, regardless of the frequency of the collocations as a whole. In Jiang (2009), the study focused on pedagogic materials for teaching collocations to Chinese learners. One of the findings regarding the learners' collocation use was that 93.6% of collocates belonged to Nation's (1990) vocabulary list of the 2000 most frequent words, i.e., of interest was the collocate frequencies, not the collocation frequencies. One study that incorporates both approaches is González Fernández and Schmitt (2015). The primary focus of this research was the link between frequency and productive collocation knowledge, which, matching Durrant (2014), was found to be weak. However, incorporating elements of a collocate frequency approach, only collocations whose constituent words were in K1-5 were analyzed to somewhat control for the effects of word frequency. Ultimately, both approaches are reasonable for selecting which collocations to teach from a frequency-based perspective (Nizonkiza & Van de Poel, 2019).

In the section on lexical sophistication (Section 2.2.2), we saw how measures like AG relied on lemma frequency data from external reference corpora, and that these measures were useful predictors of writing quality and learner proficiency. Using the collocate frequency approach, we can leverage these measures to inspect collocations and to classify them as 'low-', 'mid-', or 'high-frequency' using the thresholds described in Section 2.2.4. To my knowledge, no studies have yet to apply the 'mid-frequency' label to collocations. Using a collocation frequency approach, this type of classification would not be possible given the lower frequency numbers (which are also more variable across corpora). Although the labels 'high-frequency' and 'low-frequency' have been used in collocation frequency studies, their definitions are variable. For example, 'low-frequency' has been used to mean <5 occurrences in the BNC (Durrant & Schmitt, 2009), <6 occurrences in the BNC (Siyanova-Chanturia & Schmitt, 2008), and 5-99 occurrences in COCA (Yoon, 2016).

3.3.2 L2 collocations and association scores

Returning to collocate associations, we see that there is a relationship between sophistication and the association metrics from Section 3.2.2. Recall, MI emphasizes lower-frequency combinations, which will naturally contain more lower frequency words, and t-scores value higher frequency combinations which will contain more high-frequency words.

In studies reporting MI and t-scores, there is a general consensus that MI correlates to learner proficiency, with more advanced learners using collocations with higher MI (Bestgen & Granger, 2014; Granger & Bestgen, 2014; Naismith & Juffs, 2021; Paquot, 2019). For example, in the Bestgen and Granger (2014); Granger and Bestgen (2014) studies, the authors created a collocational profile for each text in a corpus. This profile, named 'Collgram' consists of MI scores, t-scores, and the percentage of bigrams absent from a reference corpus. The researchers calculated the Collgrams for 171 essays in the MSU corpus and 223 texts in the ICLE corpus. These same texts were assessed by expert human raters and a comparison was made. In both studies at higher proficiency levels there were lower t-scores and higher MI-scores, findings consistent with a previous study by Durrant and Schmitt (2009). Furthermore, there was a correlation between lower 'absent bigrams' and the experts' ratings, indicating that absent bigrams may be a useful proxy for collocational errors.

While promising, the conclusions from these two studies must be taken cautiously as both contained limitations. In Granger and Bestgen (2014), for texts written within one semester, MI scores did not go up as expected, possibly due to the short timeframe. In Bestgen and Granger (2014), the Collgram profiles did differentiate between proficiency levels, but the levels selected were the broad B and C bands from the CEFR which are drastically different. A more fine-grained differentiation between levels within each band would be more telling, i.e., B1, B2, C1, C2. To

date, there are minimal studies incorporating the Collgram profile and further validation is needed (Lenko-Szymanska, 2019).

Whereas there is ample evidence for using MI to assess learner text quality, the findings relating to t-scores are mixed. Some studies have found the use of collocations with high t-scores to be characteristic of learner writing; learners use more collocations with high t-scores than expert speakers (Durrant & Schmitt, 2009), and intermediate learners use more collocations with high t-scores than advanced learners (Granger & Bestgen, 2014). Other studies have not found a significant correlation between t-scores and characteristics of learner production, including essay quality (Bestgen & Granger, 2014), oral proficiency (Uchihara et al., 2021), and collocation acceptability (Naismith & Juffs, 2021). Given these mixed findings, it seems prudent to consider multiple association measures when analyzing potential collocations (Evert, 2009).

3.3.3 Collocational accuracy

With respect to accuracy, FS accuracy (including collocations) can be defined broadly as "the use of acceptable and expected multi-word units" (Crossley et al., 2013, p. 112). Collocational accuracy is especially important in academic writing as bad collocation use indicates to the reader a lack of academic expertise (Henriksen, 2013) and forces readers to decompose the collocations rather than more fluently processing them as single chunks (Howarth, 1998). Even if the meaning of the words are not obscured, collocation errors can cause 'lexical dissonance' (Hasselgren, 1994), putting a strain on the reader. 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. Of the types of lexical errors outlined in Section 2.3, collocational errors are one the most common, e.g., in Hasselgren (1994) they accounted for 19-27% of lexical errors,

and in Xie (2019) they were the most prevalent and frequent lexical error type, occurring in 96.1% of learner essays an average of 6.75 times. Some authors (e.g., Howarth, 1998; Nesselhauf, 2005) have further sub-divided collocation errors based on the reason for the error (transfer, analogy, avoidance, etc.), but these distinctions will not be considered here. Likewise, an interesting area for future research would be a consideration of defining collocation error from the perspective of categorical issues in the lexicon, e.g., in relation to transitivity with verbs.

In terms of the prevalence of collocational error rates, (Nesselhauf, 2005) found that there were collocational errors in 50% of the collocations annotated, and Laufer and Waldman (2011) calculated a 33% error rate in their data. This wide range of findings is important because it signifies that there is no established standard against which to compare new results, though the expectation is for collocational accuracy to account for significant variance in judgments of proficiency. Factors which may affect collocation accuracy rates include the approach taken to identifying potential collocations (see Section 3.2.1), the types of collocations under investigation, and the L1s of the learners (Fan, 2009; Gilquin, 2007; Nesselhauf, 2005). For example, in Fan's (2009) corpus of Hong Kong learner texts, there are multiples instances of the erroneous collocation "left/right face" which is a literal translation from the equivalent Chinese collocation.

3.4 Chapter summary

In this chapter on collocational proficiency, we have seen that collocations are a frequent and well-researched type of formulaic sequence, with properties of both lexical and grammatical structures. Overall, proficient use of collocations is a critical element of lexical proficiency due to the impact on readers/listeners. However, collocation use by L1 and L2 users differs significantly

and developing collocational proficiency is a challenge, both in terms of sophistication and accuracy. It is therefore important to identify and measure collocation use by learners, with a view to improving future pedagogical decisions. In this vein, collocation studies have developed several collocation measures for quantifying collocation use, foremost among them mutual information (MI).

4.0 Comparing human ratings and statistical measures

Chapters 2 and 3 have looked at a variety of statistical measures relating to the use of single- and multi-word lexical items, and discussed how these measures correspond to language proficiency. However, in the history of ELT, learners' spoken and written production has primarily been, and continues to be, assessed by human raters. This chapter examines the characteristics of human ratings of student production, including individual differences of raters, types of rating scales used, and a model for interpreting ratings. It then looks at the methods and findings of studies which compare statistical measures and human ratings, focusing on two studies of particular relevance to the dissertation research. Comparison studies of this nature are important for linking human ratings and statistical measures, as clearly articulated by Bulté and Housen (2014, p. 43):

> Since writing proficiency, and progress thereof, is typically measured by subjective ratings by expert evaluators, it is also important to know which features of written performance correlate with, and may influence, overall perceptions of progress by such evaluators.

4.1 Human ratings

In contrast to computer-based measurements like those listed in Section 2.2, human performance assessments of language production are inherently subjective, based on individuals' perceptions of quality. Raters are therefore central to this process, though this human element does

introduce a number of factors which can impact the validity and reliability of the assessment (Attali, 2016; Eckes, 2012; Enright & Quinlan, 2010; Griffin, 1990). As Enright and Quinlan (2010, p. 330) explain, "when scoring essays, human raters have their strengths and weaknesses, and those weaknesses, especially variability in the way raters interpret and apply rubrics, tend to undermine assessment quality." Assessing essay writing in particular places a high cognitive demand on raters (Eckes, 2012), and even looking at lexis alone, assigning a numeric score is a challenge with many elements to take into account (Fritz & Ruegg, 2013). Ratings may vary across raters (inter-rater reliability) or may 'drift' for one rater across texts (intra-rater reliability) (Wilson et al., 2017). And yet, this type of assessment is still widely administered because it directly tests communicative language ability (Hamp-Lyons, 1990; Hawkey & Barker, 2004). As a result, it is thought to have worthwhile high validity, even at the cost of lower reliability (Agustín Llach, 2011). Efforts are therefore made to mitigate the intrinsic subjectivity of rating through various means such as carrying out rater training, double marking texts, and using rating scales.

4.1.1 Rater characteristics

Considering the behaviours of raters, numerous studies show that raters differ in their degree of severity/leniency (Eckes, 2012; Lumley & McNamara, 1995; McNamara & Adams, 1991/1994; Schaefer, 2008; Wilson et al., 2017). For example, in McNamara and Adams (1991/1994), the researchers found extensive variability in the severity of four IELTS writing examiners. Based on their Many-Facet Rasch Measurement (MFRM) models (to be discussed in Section 4.3), if the likelihood of the most lenient rater giving a certain score was 50%, the chance that the harshest rater would give that same score was only 12%.

This variability may be partially accounted for by the unevenly perceived importance of different criteria (Eckes, 2012; Goh & Ang-Aw, 2018; Zhang & Elder, 2014) or differences in how rubrics are interpreted (Vaughan, 1991). If raters consider certain criteria to be more important, it can lead them to focus more on those criteria's features, such as relevance of topic, text length, or grammar (Hall & Sheyholislami, 2013; Orr, 2002) and can colour their interpretation of rating scales (Lumley & McNamara, 1995). There are also commonly observed patterns including the central tendency effect in which some raters avoid the extremes of rating scales, and the halo effect in which strengths or weaknesses in one analytic category unintentionally affect raters' views for other categories (Eckes, 2009; McNamara et al., 2019; Wilson et al., 2017).

To understand these behaviors, studies have examined raters' thought processes by using a think aloud protocol during the rating process (see Barkaoui, 2011 for a review of these studies). In such studies, there is some evidence that lexis is not of primary consideration. For instance, in Lumley and McNamara (1995) lexis was only mentioned 12.3% of time, leading the authors to conclude that lexis was not a major component of text evaluations. Similarly, in Cumming et al. (2002) mentions of lexis were not prevalent. However, using think aloud protocols has been shown to actually alter the thought process of the raters (Barkaoui, 2011; Lumley, 2005), so any conclusions in this regard must be considered tentative.

Goh and Ang-Aw (2018) revealed the potential for disparity between concurrent and retrospective comments. In this study, the researchers elicited teachers' thoughts during an oral assessment task and also one month later with a questionnaire. Although the questionnaire indicated that vocabulary range was important for all the teachers, this belief was not reflected in the comments during the rating process. A few other studies have also elicited raters' thoughts after the rating process is complete (e.g., Connor-Linton, 1995; Hall & Sheyholislami, 2013;

Lenko-Szymanska, 2019). In general though, raters' beliefs have been understudied (Goh & Ang-Aw, 2018), and even less is known about how raters arrive at their decisions with regards to lexical scores (Fritz & Ruegg, 2013).

In contrast, raters' past experience has received more attention, including personal background, professional training, and work experience (Pula & Huot, 1993). Past experience is more easily quantifiable than rater cognition, and it is also highly valued in the teaching community. Experience may take the form of past ELT experience or rating experience. Song and Caruso (1996) found that ELT experience led to significant holistic scoring differences, but not analytic scoring differences when comparing the ratings of ELT and English faculty members. Qualitative studies have also shown that the thought processes of experienced raters differ from novice raters (Barkaoui, 2011; Cumming, 1990; Weigle, 1994). However, other quantitative studies did not see a significant improvement in rating accuracy for more experienced raters (Brown, 1991; Lim, 2011; Shohamy et al., 1992; Weigle, 1998). As a result, the extent to which past experience affects ratings remains unclear though intuitively one might expect such experience to improve rating reliability.

One final element of rater characteristic that has been studied is linguistic background. In general, the rater's first language has not been found to impact scores (Connor-Linton, 1995; Johnson & Lim, 2009).

4.1.2 Rater training

Assessment training is one means of developing rater expertise, with the specific goal of increasing consistency of scoring (Attali, 2016). Unlike with rater experience, the efficacy of rater training is more clear-cut. Rater training has been shown to improve intra-rater reliability (Brown,

2006; Hall & Sheyholislami, 2013; McNamara, 1996; Weigle, 1994, 1998) and adherence to rubrics (Davis, 2016), which in turn leads to more reliable results. For example, in Linnarud (1986), three groups rated L2 English student essays. The group of L1 Swedish teachers who had received assessment training showed high agreement in their ratings whereas the two groups without the assessment training (L1 English teachers and L1 English non-teachers) showed low agreement. In a more extreme finding, Attali (2016) showed that after only brief training, inexperienced raters produced results comparable to highly experienced raters using a six-point holistic scale to rate essays, though there was more in-group variance. Of course, as Wilson et al. (2017, p. 17) remind us, "humans are not machines and assessing writing is not easy"; even with training, we cannot expect all biases and challenges to be eliminated.

4.1.3 Scale types

Turning to the types of scales used for assessing writing, we find that they are typically divided into two categories: holistic and analytic. These scales are the most common way to assess L2 writing composition tests (Hall & Sheyholislami, 2013), and they usually include a vocabulary component (Agustín Llach, 2011). Holistic scales assess the overall quality of an essay, whereas analytic scales assign scores for individual aspects such as topic relevance, grammar, and vocabulary. For example, TOEFL uses a holistic five-point scale as part of its assessment of their independent writing task (TOEFL, 2019). In contrast, IELTS uses only analytic scales whose scores are then averaged to produce a single overall score from 1 to 9 (IELTS, n.d.-c).

Much as these two international exams differ in their choice of scale type, so too do SLA researchers. On the one hand, some researchers have opted to have human raters assess learner writing using only holistic scales (e.g., Enright & Quinlan, 2010; Guo et al., 2013; Jarvis, 2013a;

Kyle & Crossley, 2015). Other researchers prefer for raters to use only analytic scales (Aryadoust & Liu, 2015; Dabbagh & Janebi Enayat, 2019; Fritz & Ruegg, 2013). A third popular option is to use a combination of holistic and analytic scales (e.g., Crossley et al., 2015; Granger & Bestgen, 2014; Lenko-Szymanska, 2019; Vögelin et al., 2019). Typically, holistic scales are accompanied by descriptors, i.e., prose describing what each point on the scale represents. Of the studies cited above, only Jarvis (2013a) does not provide descriptors to avoid introducing potential rater bias. However, this methodological decision led to poor inter-rater reliability.

The preference by many researchers to use both holistic and analytic scales may stem from the fact that both scale types inherently possess strengths and weaknesses. Holistic scales are undoubtedly more efficient as they can be quickly implemented and the results more easily interpreted (Wilson et al., 2017). Holistic scales may also be considered to have high validity as they more closely match actual reading. Typically in authentic situations, readers form a single overall impression of the quality of a text (Weigle, 2002) which is "greater than the sum of the text's countable elements" (Hamp-Lyons, 1990, p. 79).

However, on their own, holistic scores are not very informative (Wilson et al., 2017) and oversimplify the construct of writing proficiency (Cumming et al., 2002). Learners often have 'jagged' profiles in which not all dimensions are equally proficient (Eckes et al., 2016; Weigle, 2002), and raters tend to weight the dimensions of grammar and vocabulary more heavily compared to content and organization (Cumming et al., 2002). A reliance on holistic scales alone therefore positively biases candidates whose strengths are grammar and lexis. Holistic ratings have also been shown to be unreliable across studies (Hamp-Lyons, 1990). It is perhaps for these reasons that TOEFL switched policies, citing challenges with consistency, reliability, and efficiency (Enright & Quinlan, 2010). Whereas previously two humans holistically rated tasks, currently

there is a combination of one holistic human rating and with automated ratings of analytic categories.

In contrast to holistic scales, analytic scales allow for raters to assess writing more systematically as the scales themselves are more detailed and multi-faceted (Weigle, 2002). Thus, learners with jagged profiles can receive different ratings for different writing dimensions. Consequently, analytic ratings can also better inform writing feedback and instruction based on specific strengths and weaknesses of a text. Despite this additional level of detail, analytic descriptors may still be 'fuzzy' as dimensions of language are not always easily separable (Vögelin et al., 2019). For example, multi-word lexical units could be considered part of either the systems of grammar or vocabulary depending on one's theory of language, as discussed in Section 3.2.3. It is also certainly the case that using analytic scales is more time consuming than holistic scales (Wilson et al., 2017).

4.2 Comparison studies

As previously noted, there is a strong rationale for conducting comparison studies (Bulté & Housen, 2014). This endeavor is a clear one, though the results are anything but, in part because of the high number of text features correlating to ratings (Agustín Llach, 2011). To further complicate matters, studies in this field vary greatly in terms of task parameters and formats (Crossley et al., 2014), as well as other methodological elements.

Comparison studies have compared ratings of both speaking (e.g., Crossley et al., 2011; Lu, 2012; Yoon et al., 2012) and writing (Enright & Quinlan, 2010; Monteiro et al., 2020; Vögelin et al., 2019), as well as different aspects of language systems and skills, e.g. formulaicity (Ellis &

Simpson-Vlach, 2009) and syntactic complexity and cohesion (Guo et al., 2013). More than any other textual aspect, lexical features have been the focus of these comparisons (Vögelin et al., 2019). As early as the 1980s, Arnaud (1984) and Linnarud (1986) examined textual measures of lexical diversity, sophistication, and accuracy. In the case of Arnaud (1984), the measures were compared to students' vocabulary test scores. In Linnarud (1986), these measures were compared to a native speaker comparison group. Since the time of these two studies, our understanding of lexical constructs has progressed significantly as have the methods and tools for extracting the relevant information.

4.2.1 Comparison study methods

In this section, the focus is on the most crucial components of comparison studies: the raters, the learners, the texts, the assessment scales, and the ensuing statistical analyses. Appendix A presents a tabular overview of 27 such studies as a sample, not a complete meta-analysis. Here, studies were included focusing on L2 writing (unlike, e.g., Riemenschneider et al., 2021), and on only the L2 groups when an L1 control group was present. In addition, only studies, or the portions of studies, considering independent writing tasks were analyzed, and not integrated writing tasks (as in Guo et al., 2013). Finally, only comparison studies are considered that compare statistical measures to proficiency ratings, not studies comparing statistical measures to other proficiency metrics like grade point average (e.g., Morris & Cobb, 2004).

4.2.1.1 Raters

Raters are often described in terms of their expertise, e.g., they are labelled 'expert raters'. However, what this term actually means is often not explained or falls short of the definition of

'expert rater' by international assessment standards. Some previous studies have used raters with limited or uncertain expertise (e.g., Crossley et al., 2015; Ruegg et al., 2011; Vögelin et al., 2019), though others have used more experienced raters (e.g., Guo et al., 2013; Lenko-Szymanska, 2019). We can also observe that the number of raters spans a broad spectrum, with two raters the most common choice (e.g., Dabbagh & Janebi Enayat, 2019; Grant & Ginther, 2000), but extending up to as many as 45 (Ruegg et al., 2011). Correspondingly, the number of ratings for each text is largely dependent on the number of raters. Most commonly, texts are also rated twice, with a third rater sometimes adjudicating divergent cases (Agustín Llach, 2007; Jiang et al., 2021; Kyle & Crossley, 2016). Although this practice helps to ensure interrater reliability, it does not account for issues which may affect both raters, like severity in rating one analytic category. One exception is Vögelin et al. (2019); although not explicitly stated, based on the numbers of raters (37), texts (8) and ratings per rater (4), it appears that each text was rated an average of 18.5 times.

4.2.1.2 Texts

The number of learner texts analyzed is fairly consistent across studies, with most studies looking at between 200-300 (Granger & Bestgen, 2014; Guo et al., 2013; Jarvis, 2013a; Kyle & Crossley, 2015; Lenko-Szymanska, 2019; Yu, 2010). Earlier studies tended to rate fewer texts, e.g. Engber (1995) with 66 texts and Linnarud (1986) with 63. More recently, Ruegg et al. (2011) considered 140 texts, and Vögelin et al. (2019) looked at only 8 texts. Importantly, the Ruegg et al. and Vögelin et al. studies are those with the highest number of raters, though no statistical justification is given for this choice. These two recent outliers point to a common trade-off between the number of raters, the number of texts, and the expertise of raters. Studies either use a small number of expert raters and a large number of texts (e.g., Guo et al., 2013) or a large number of inexpert raters and small number of texts (e.g., Vögelin et al., 2019).

The genre of the texts for all of these studies fall under the broad umbrella of ‘essays’. Often learners are asked to give an opinion on a subject, in which case the essays are described as an ‘expository essay’ (also refered to as ‘argumentative’ essay , e.g., Fritz & Ruegg, 2013; Kyle et al., 2020). The other commonly analyzed essay type is the descriptive essay (e.g., Dabbagh & Janebi Enayat, 2019; Daller & Phelan, 2007). Within these genres, studies often try to control for task effects by limiting participants to one prompt (e.g., Ginther & Grant, 1997; Lenko-Szymanska, 2019); when learners have a choice of tasks, it lowers the reliability of scoring by introducing measurement error (Polio & Glew, 1996). However, not all studies restrict writers to one prompt; the Bestgen and Granger (2014); Granger and Bestgen (2014) studies used eight different prompts. The variable ‘time on task’ is also typically controlled for by setting a writing time limit. Most often, this time is either 30 or 35 minutes (e.g., Ferris, 1994; Monteiro et al., 2020).

4.2.1.3 Learners

The overall number of learners is typically the same as the number of texts, with each learner producing one text. Exceptions include studies in which learners produced two texts each (Aryadoust & Liu, 2015; Bulté & Housen, 2014; Crossley et al., 2014), three texts each (Bestgen & Granger, 2014), or in the case of Kyle and Crossley (2015), 10 learners produced 180 texts (not evenly distributed). In nearly all the modern comparison studies reviewed, the texts are written by learners from two or three proficiency levels. Normally, these levels are intermediate or higher, which makes sense given the proficiency required to write essays, e.g. upper-intermediate and advanced (Lenko-Szymanska, 2019); intermediate and advanced (Granger & Bestgen, 2014); intermediate to advanced (Ruegg et al., 2011); ‘High’ and ‘Low’ (Vögelin et al., 2019); and ‘High’, ‘Medium’, and ‘Low’ (Monteiro et al., 2020). The learners most characteristically have varied L1s

(e.g., Daller & Phelan, 2007; Ferris, 1994), though a number of studies also focus on L1 Chinese learners exclusively (Aryadoust & Liu, 2015; Crossley et al., 2012). All of the studies reviewed focused on L2 English writing.

4.2.1.4 Scales

As discussed in Section 4.1.3, researchers have opted for either holistic scales, analytic scales, or a combination of both. In the sample, these three categories represented 15, 7, and 5 studies respectively. However, the specific scales selected in these studies differs greatly, though they might be categorized as either 'standardized' or 'bespoke'. Standardized scales are those that are publicly accessible and widely used for assessment purposes (e.g., Guo et al., 2013; Vögelin et al., 2019). The advantages of such scales are their face validity, their accessibility, and the rigorous review and revision which they have undergone, often over many years. In contrast, bespoke scales are created by the researchers. The purpose in doing so is to better target the constructs under scrutiny (e.g., Crossley et al., 2012; Crossley et al., 2013) or because they align with the assessment system used in the context of the corpus (e.g., Granger & Bestgen, 2014).

4.2.1.5 Statistical analyses

To unearth the relationship between lexical metrics and human judgments, researchers rely on a range of statistical tests. Often, correlations are calculated with Pearson correlations (e.g., Bestgen & Granger, 2014; Guo et al., 2013; Jarvis, 2013a), though other types are seen as well, e.g. Spearman rank-order correlations (e.g., Lenko-Szymanska, 2019), Spearman rho correlations (e.g., Daller & Phelan, 2007; Lu, 2012), and ANOVAs (e.g., Vögelin et al., 2019). In most recent

studies, regressions are also used to the predict human judgments of test data (e.g., Crossley et al., 2012; Crossley et al., 2013; Monteiro et al., 2020; Ruegg et al., 2011; Yu, 2010).

With respect to the use of writing scales, each scale point is typically accompanied by a descriptor. In research using writing scales, these instruments are considered to have interval scale properties. This position is certainly not uncontroversial as there is current disagreement on the matter (Amidei et al., 2019; van Rijn, 2019). For practical reasons, it can be argued that whether or not test scores are true interval scales is not relevant given that there are standardized norm- and criterion-referenced interpretations, and because there are many examples in different fields where scales with unequal units are used (Van der Linden, 2015/2016 cited in van Rijn, 2019). Schütze and Sprouse (2013) likewise endorse this practice if the judgment data is treated appropriately. This statistical choice also appears to be a common practice. In a methodological meta-analysis of 135 papers, Amidei et al. (2019) reviewed the use of Likert and rating scales in the field of Natural Language Generation. They found such scales often produced data which were then used in parametric tests as though the data were interval scale measurements.

4.2.2 Comparison study findings

Viewing the body of research as a whole, certain trends emerge. One important finding is that text length is a key feature, with longer texts receiving higher ratings (Ferris, 1994; Grobe, 1981; Guo et al., 2013; Linnarud, 1986). As Kyle et al. (2020) note, this relationship may in part exist because length can be considered evidence of idea generation. In Guo et al. (2013), of the five significant variables, text length accounted for an incredible 47.8% of variance, with all other significant factors under 1%. Likewise, Ferris (1994) found that 37.6% of variance between levels was accounted for by text length, with higher rated texts containing more words. This finding is

important because in models of language proficiency, text length can 'wash out' the predictive strength of other variables (Crossley & McNamara, 2012). Whenever possible, it is therefore advisable that text length be controlled for in order to better understand the effect of other variables.

With regards to lexis, Crossley et al. (2013) found that automated indices related to lexical breadth (including sophistication and diversity) correlated more strongly than indices related to lexical depth. This conclusion matches other studies focusing on sophistication and diversity. For example, raters assign higher scores to essays with lower-frequency words (Crossley & McNamara, 2012), which can be captured as part of a variety of sophistication metrics (Bulté & Housen, 2014; Daller & Xue, 2007). In fact, in the Lenko-Szymanska (2019) study, although the entire model only accounted for 35% of variance between the two proficiency groups, frequency-based diversity and sophistication measures showed a strong correlation with human judgments. Guo et al. (2013) also discovered four significant variables which fall under the umbrella of lexical sophistication. Interestingly, Vögelin et al. (2019) likewise found an interaction between lexis scores from raters and texts with higher sophistication; however, these higher analytic ratings did not extend to the holistic scores. Yu (2010) focused on diversity rather than sophistication in their study (operationalized as D), and metric matched human raters' proficiency ratings, explaining 11% of variance. As reviewed earlier, lexical accuracy has shown to be a significant variable (Polio & Shea, 2014; Ruegg et al., 2011), as has collocational proficiency (Bestgen & Granger, 2014; Crossley et al., 2015).

4.3 Many-facet Rasch measurement

It is tempting to picture the score given to a written essay as a true reflection of the writer's ability. However, no matter what steps are taken to increase reliability, the final rating is invariably a reflection of multiple factors, including the writer's language proficiency (grammatical, lexical, etc.), the characteristics of the rater, and the writing prompt (Eckes, 2009; McNamara et al., 2019). That is not to say that reliability cannot be improved; as discussed, best practice in assessment often includes rater training, the use of scales, and multiple ratings per text.

Even then, human raters do not always agree, and for a number of reasons. For example, individual raters may adjust their scores depending on the perceived difficulty of an item (Polio & Glew, 1996). As we have seen, across raters there are also typically differences in the degree of leniency/severity, interpretation of scales, and levels of consistency. Overall, there is substantial evidence that there is systematic error in human ratings. This 'luck of the draw' when it comes to raters is an unwanted variable, threatening the validity of writing assessments (Eckes, 2009). For this reason, it is essential to consider rater effects from a statistical perspective (Robitzsch & Steinfeld, 2018) using tools such as many-facet Rasch measurement models (MFRM; Linacre, 1989, 1994).

In essence, MFRM models are a subset of Rasch measurement models which probabilistically "predict the outcome of encounters between persons and assessment/survey items" (Aryadoust et al., 2021, p. 7). However, whereas other earlier forms of Rasch measurement are dichotomous, MFRM allow for simultaneous consideration of multiple variables. These variables are referred to as *facets*, including any component which might systematically affect a test score (Eckes, 2009) such as the rater, the candidate, and the test item. Essay writing can therefore be considered rater-mediated assessment as it includes a rater facet (Eckes, 2009).

MFRM models are considered ideal for investigating assessment of production and as a validation tool because they compensate for measurement error (Eckes, 2009). At present, MFRM models are widely used in language assessment research, especially in relation to writing (Aryadoust et al., 2021; Eckes, 2015; Lumley, 2005; McNamara, 1991, 1996).

In MFRM models, each person's ability, rater's severity, or item's difficulty is viewed as a fixed model parameter with distinct attributes (Eckes, 2009; Robitzsch & Steinfeld, 2018). Items might be individual test questions (e.g., different writing prompts) or the analytic bands used to assess production (grammar, lexis, etc.). In a common MFRM analysis, the writer, rater, and items are the three facets and can be conceptualized as independent variables influencing the rating scores. In the parlance of MFRM, these facets are *proximal factors* as they have an immediate impact on scores. Other possible facets such as demographic information about raters or examiners are called *distal factors* as their potential impact is less direct. The output of this model would then be the rating, expressed in log-odds units (logits) and accompanied by standard error, with the intention of compensating for differences in rater severity and item difficulty (Eckes, 2009). The exact formula and description from Eckes (2009, p. 13) is as follows:

$$\ln\left[\frac{p_{nijk}}{p_{nijk-1}}\right] = \theta_n - \beta_i - \alpha_j - \tau_k$$

p_{nijk} = probability of examinee n receiving a rating of k on criterion i from rater j
p_{nijk-1} = probability of examinee n receiving a rating of $k - 1$ on criterion i from rater j
θ_n = proficiency of examinee n
β_i = difficulty of criterion i
α_j = severity of rater j
τ_k = difficulty of receiving a rating of k relative to a rating of $k - 1$

Usefully, logits can then be transformed into *fair scores* (or *fair averages*) which report the same scores on the raw-score scale. In other words, a fair score is the score that would be expected for

that item from a rater with average severity (Linacre, 2021). If fair scores are used, then rater severity ceases to be a significant issue, increasing the validity of interpretations of the resulting scores (McNamara et al., 2019). Typically MFRM models are created using one of two popular programs (McNamara et al., 2019): WINSTEPS (Linacre, 2021) and Facets (Linacre, 2020). These same models can also be created using R (e.g., Robitzsch & Steinfeld, 2018).

How MFRM models are used varies across studies depending on the focus (Robitzsch & Steinfeld, 2018). On an international level, the Council of Europe (Council of Europe, 2001) used MFRM as the methodological basis for the CEFR descriptor scales (Eckes, 2009). Assessment bodies like IELTS and Cambridge Assessment also use MFRM models as part of their preparation of training materials (Griffin, 1990). However, in the comparison studies reviewed, MFRM do not figure prominently and are only used in on study, Aryadoust and Liu (2015). In this study, the writing ability of Chinese EFL learners was analyzed, comparing statistical measures of mental representation from Coh-Metrix (Graesser et al., 2004) against analytic ratings from four raters. Prior to carrying out these comparisons, the authors used MFRM models to check for any effects of rater severity and to adjust the scores if necessary, i.e., using fair scores. This technique of using MFRM to validate collected data prior to conducting further statistical operations is an effective method for increasing the reliability of any subsequent conclusions.

4.4 Key studies

With respect to the dissertation research presented in Chapters 5 and 6, there are two comparison studies which are of particular relevance based on their methodologies and research focuses, Fritz and Ruegg (2013) and Read and Nation (2006).

4.4.1 Fritz and Ruegg (2013)

The majority of the comparison studies reviewed thus far have used corpus methods to compile the texts to be assessed. Fritz and Ruegg (2013) presents an interesting outlier. These researchers used an experimental design to investigate the extent to which raters are sensitive to different lexical qualities in learners' writing. Like other studies, they focused on argumentative essays written under timed conditions (30 minutes), responding to one prompt ("the merits of vegetarianism"). The learners were L1 Japanese/L2 English speakers studying at a university in Japan. 27 experienced raters used four analytic scales to assess the essays, and these ratings were analyzed using ANOVAs to find the relationships to measures of accuracy, diversity (called 'range' in the study), and sophistication. The findings indicated that lexical accuracy significantly predicted ratings, though diversity and sophistication surprisingly did not.

Of more interest here is the methodology employed for creating the instruments. Rather than analyzing a wide range of essays, a single 'base' essay was used. The 32 content words in this text were manipulated to create 27 total versions: low/mid/high versions of accuracy, diversity, and sophistication, i.e. a 3x3 design. To manipulate accuracy, part-of-speech and meaning errors were introduced (low accuracy = 32, mid accuracy = 16, low accuracy = 0). For diversity, the number of unique content words were altered (low diversity = 18, mid diversity = 25, high diversity = 32). For sophistication, all 32 content words were placed in specific frequency bands (low sophistication = K1, mid sophistication = K2-3, high sophistication = K4+).

Through the use of this methodology, the study tightly controlled all other text variables other than the targeted 32 lexical items, including length, topic, cohesion, and grammar. As a result, the findings could be clearly interpreted, and the study itself could be replicated or adapted for future research. However, as the authors themselves acknowledge, there were certain

limitations. Critically, each rater assessed three of the 27 versions. Because each of these versions would appear very similar, they were mixed in with other authentic texts that the raters were also assessing as part of their teaching duties; in total each rater assessed 39 texts, 3 of which were from this experiment. However, in the results, some of the manipulated texts were awarded a score of 0, indicating that perhaps the raters believed the texts to be too similar to one another and the result of cheating. A related issue in this regard is the naturalness of the manipulated texts. All 27 versions were based one on low-level text, and the manipulations were confined to 32 specific lexical words. As a result, there was not always a wide range of synonyms, and the most advanced version provided in the study's appendices stands out as potentially inauthentic, with unclear use of lexis, e.g.,

> Second reason is about my health. I have comprehended that meats need our mortal torso because meats fortify our hemoglobin and support our anatomy, so only vegetable is not advantageous for our health. (manipulated words underlined)

Another potential shortcoming is the operationalization of sophistication. The authors considered lexis in K4+ to be advanced, though as we have seen in Section 2.2.3, this is lower than the K9 or K10 threshold more commonly used by researchers based on coverage percentage. The lack of significant findings with respect to sophistication may therefore be a result of too narrow a range of frequencies. From this study, the dissertation study in Chapter 6 will therefore adopt many of these same general approaches to experimental control through the use of carefully manipulated versions. However, only one version of each text will be rated by each rater, the naturalness of the texts will be prioritized, and a wider range of frequency bands will be used to focus on sophistication.

4.4.2 Read and Nation (2006)

Read and Nation (2006) was not included in the discussion of comparison studies as it focuses on learner speech, not writing. It is nevertheless important to consider as it overlaps to great extent with the focus of this dissertation. In their study, Read and Nation investigated the vocabulary use of IELTS test takers. For the data, the study analyzed 88 recordings from various exam centers of learners completing their Part 2 'long turns', i.e., monologues based on a prompt. These 88 recordings were awarded ratings between 4 and 8 on the IELTS 9-point scale, equivalent to low B1 to high C1 on the CEFR. The transcribed texts were then analyzed using software to calculate metrics of lexical diversity (vocD) and lexical sophistication (LFP profiles), and the texts were analyzed qualitatively as well to uncover trends in vocabulary use.

The results indicated a great deal of individual variation within levels. However, in general, speech with higher ratings had higher lexical diversity and a higher percentage of low-frequency vocabulary. The qualitative analysis also uncovered that at the highest levels, speech was characterized by "mastery of colloquial or idiomatic expressions" (Read & Nation, 2006, p. 22), not just individual words. There was no quantitative measurement of the use of FSs, pointing to the need for further research of this element of lexical IELTS ratings. Methodologically, this study does not closely correspond to that of the dissertation research. Where the similarities lie is in the raters and scales (IELTS-based) and in the proficiency bands (to be discussed in Chapter 5). As well, one of the desired potential outcomes is much the same, only in relation to writing: to inform lexical rating descriptors which will make salient the most impactful lexical features.

4.5 Chapter summary

In this chapter we have seen that there is a long history of experts assessing learners' writing compositions, and that this form of assessment continues to hold perceived high validity. However, the use of raters necessarily introduces a subjective element to the assessment process; there are a number of individual differences which must be considered, with rater training foremost among them. In addition to training, rating scales are usually used to maximize objectivity, based on holistic and analytic descriptors. Even with these steps, interrater reliability remains a thorny issue, which is why models like MFRM have been developed to account for rater differences. Taken together, these assumptions suggest that studies involving very limited numbers of raters are in danger of being less reliable because rater differences will have a greater effect on the results. Therefore, when possible, a larger pool of expert raters is preferable. For the dissertation research it is therefore essential to consider the methodological choices with respect to scales and raters, prioritizing rater expertise and the use of both holistic and analytic scales. Conducting statistical analyses comparable to those of previous studies, including MFRM and linear regressions, also allows for greater comparison to existing research in this area and future replication.

5.0 Validation of text length normalization

We now move from the literature review to a preliminary study in which instruments for the main study are created and validated. As seen in Chapter 4, to maximize the validity of comparison studies, the texts being rated should ideally be of equal length so that the effects of text length do not mask other factors affecting proficiency ratings (Crossley & McNamara, 2012). This chapter presents a set of three Academic IELTS Task 2 essays which can be used for a comparison study, but which are of different lengths. First, the pertinent characteristics of the texts are described, including the source, text type, and other factors. The lengths of the three texts are then normalized using precise manipulations, and these two versions ('original' and 'normalized') are compared using numerous statistical measures. Next, 'base' texts are created from the normalized versions by adjusting the number of accurate and inaccurate collocations in each to stratify the texts' collocational profiles. Finally, experts' ratings of all three text versions (original, normalized, and base) are analyzed to see whether the overall text quality of each version remains unchanged. If there are no significant differences in terms of CEFR level, the base texts can be used for the comparison study in Chapter 6 as representative samples of different levels. However, if there are significant differences, then the proficiency variable cannot be reliably included in the ensuing models.

5.1 Original texts

There are three original texts, all from the same source, the academic writing sample texts from the IELTS website (Appendix B; IELTS, n.d.-d, pp. 11-15, Academic Writing sample materials and examiner comments). IELTS has a longstanding history as an international exam, founded in 1989 as a joint project between British and Australian governmental agencies (McNamara et al., 2019). Each year it is taken by over 3.5 million candidates in 140 countries for purposes of study, work, and migration (IELTS, 2019).[8] The exam also has high perceived validity from teachers and students (Coleman et al., 2003) who have a positive attitude towards the exam (Green, 2019).

As such, the exam carries high stakes and carries great weight in the field of ESL/EFL, both for language learners (Rea-Dickins et al., 2007) and language teachers (Estaji & Ghiasvand, 2019). There is a substantial body of existing literature about IELTS (often funded by IELTS itself). Overall, this work has shown IELTS to be predictive of academic success in a number of contexts (e.g., Hill et al., 1999; Oliver et al., 2012; Schoepp, 2018), with most studies indicating a positive medium relationship (Dang & Dang, 2021). For more critical debate about aspects of IELTS' validity, see Green (2019) and Pearson (2019). Importantly, research on IELTS has shown candidates' IELTS ratings to correlate to vocabulary size (Harrington, 2018; Milton & Alexiou, 2009).

IELTS is available in two formats: Academic and General Training, differing in content and task type. Henceforward, we will be considering only the Academic test which is the most

[8] The most recent public information is from 2019, so it is not known the extent to which Covid-19 has affected the number of IELTS test-takers in 2020/2021.

common format (76%; IELTS, 2018) and the most relevant in the context of this research. In brief, the exam consists of four parts, one for each of the skills of Listening, Reading, Speaking, and Writing, and it takes 2 hours and 45 minutes to complete (IELTS, n.d.-e)

The writing component of IELTS consists of two tasks, a letter and an essay, though only the latter will be considered here. The Task 2 essay is an independent writing task which is designed to elicit a test taker's underlying writing ability, with the writer relying only on their existing knowledge and experience (Crossley et al., 2014; Guo et al., 2013). It is also considered to be a complex performance task, appropriate for assessment because of its authenticity as a real-world task used for communicative purposes (Enright & Quinlan, 2010). Standardized tasks such as these are carried out under controlled exam conditions: there are strict time controls and writers do not have access to outside resources. In part, this may be why this type of assessment is the best researched (Weigle, 2002) and why direct testing of writing is considered to have high face, construct, and content validity (Hamp-Lyons, 1990).

The three example texts provided by IELTS are authentic texts, written under exam conditions. Therefore, the writers were instructed to spend 40 minutes writing at least 250 words in response to the following prompt (IELTS, n.d.-d, Academic Writing sample task 2A):

> *Children who are brought up in families that do not have large amounts of money are better prepared to deal with the problems of adult life than children brought up by wealthy parents. To what extent do you agree or disagree with this opinion? Give reasons for your answer and include any relevant examples from your own knowledge or experience.*

Candidates' choice of lexis on IELTS is influenced by topic (Read & Nation, 2006), so uniformity in this regard was considered essential for this experiment, even though it limited the pool of potential text choices from the public IELTS sample responses. Where the texts differ is

in the ratings awarded to them. Because they are public samples, the texts have already been assessed by IELTS and are used as models of different band scores; the reliability of these scores as a baseline is therefore ensured and they can be used as the gold standard. Helpfully, these scores are accompanied by examiner comments justifying the scores. 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 three scores correspond to the CEFR levels of B1, B2, and C1 respectively (Figure 3; IELTS, n.d.-a, p. 1). In more descriptive terms, IELTS considers a Band 4 to be a 'Limited user', Band 6 to be a 'Competent user', and Band 8 to be a 'Very Good user'. In the context of a higher education environment, guidance from IELTS suggests that a score of 6.5 or 7 is 'clearly acceptable' for students to enter 'linguistically less demanding academic courses' like mathematics (Green, 2019). Appendix C provides a more complete overall description and comparison of these levels from both the IELTS and CEFR official documentation. This alignment between IELTS band scores and CEFR levels is important for cross study comparisons that have likewise used CEFR bands to distinguish between proficiency levels (e.g., Chen & Baker, 2016; Granger & Bestgen, 2014).

The CEFR labels also add to the perceived validity of the assessments due to their widespread adoption and emphasis on language for communicative purposes (Hawkey & Barker, 2004). Although the CEFR is now skills-based, it originally contained vocabulary lists (Milton & Alexiou, 2009), and lexis remains an important component of the descriptors. In an investigation of CEFR levels and vocabulary knowledge, Milton and Alexiou (2009) found that learners required receptive knowledge of approximately 3000 words to move from the A2 to B1 band and knowledge of an additional 500 words to move to B2.

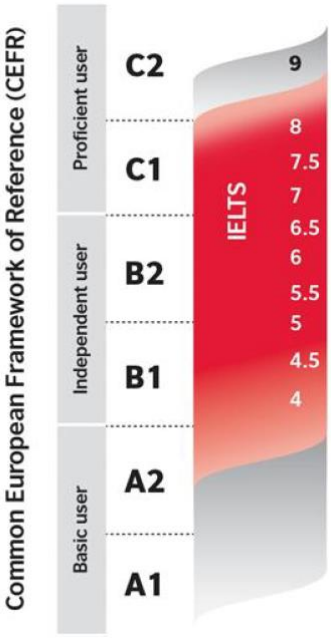

Figure 3 Comparing IELTS band scores and CEFR levels (IELTS, n.d.-a, p. 1)

Although originally handwritten, the texts have been typed for practicality and standardization purposes; research has shown that there is no significant difference between scores from paper-pencil and online IELTS tests (Chan et al., 2017). There is no background information on the authors of these texts, such as first language, age, or gender. For this study, these factors are considered here to be construct-irrelevant, as in Vögelin et al. (2019).

5.2 Normalized texts

5.2.1 Normalization process

As described in Section 4.2.2, text length may account for the greatest amount of variance in proficiency judgments and may mask the effects of other factors. To avoid this issue, it is

therefore necessary to normalize the lengths of the three original texts (normalized texts in Appendix D). Although seemingly simple, this is in fact a delicate process. The easiest way to shorten texts is through segmentation or truncation. For example, Crossley et al. (2012) and Engber (1995) selected text segments of approximately 150 and 126 words respectively from the middle of texts, and Vögelin et al. (2019) truncated texts to approximately 460 words by cutting off text beyond this threshold. This approach is practical for conducting automated textual analyses which do not consider the text as a whole. However, such an approach cannot be used if the texts are to be rated by humans as aspects such as task completion and organization will invariably be affected. Instead, the three texts were manually altered through addition, subtraction, and alteration of words as minimally required. Using my own extensive background as in IELTS examiner, examiner trainer, and teacher, I endeavoured to maintain the same levels as in the original text for each of the IELTS analytic categories.

The original text lengths are 172 (B1), 378 (B2), and 254 (C1) words. As the task expectation is at least 250 words, these counts signify that the B1 text is quite under-length, the B2 text is longer than necessary, and the C1 text is about the expected length. To increase the length of the B1 text, phrases, clauses, and sentences were added using much of the same language found elsewhere in the text. For example, the author writes *"I start work from 20 ages."* (para. 4). Later in the response, similar language is used in a new clause: *"They start when they 15"* (para. 4). It was decided not to repeat the exact erroneous structure (*"They start work from 15 ages"*) as this might inadvertently increase the salience of this error type which only occurred once in the original text. Conversely, to reduce the length of the B2 text, phrases, clauses, and sentences were removed, e.g., "*Whenever they want something, the money is easily gave to them ~~as if every day are their birthday~~."* (para. 2). Where possible, entire phrasal units were removed to maintain the

original voice of the author. From the C1 text, only four words were removed to achieve the 250-word target.

Throughout the process of text manipulation, precise quantitative analysis of the original and normalized texts was carried out to ensure that key metrics were not being significantly impacted. These measures include the following lexical, syntactic, and collocational indices that have been used in other studies to distinguish between proficiency levels:

1. *Text length*: number of tokens, counted manually using the IELTS guidelines for what constitutes a word. Stretches of language from the prompt were not counted, and contracted forms were separated for textual analysis (as in Read & Nation, 2006).
2. *Lexical diversity*: operationalized as vocD (see Section 2.2.1) using lemmas as the counting unit (see Section 2.2.4).
3. *Lexical sophistication*: operationalized as AG (see Section 2.2.2) using the Pitt Service List Level 3 (PSL3; Naismith et al., 2018) as the basis for what counts as an advanced lemma.
4. *Contextual diversity*: calculated using TAALES software (Kyle & Crossley, 2015) and operationalized as bigram range of the texts in the academic domain of COCA. Range is predictive of proficiency, especially bigram/trigram associations (Kim et al., 2018)
5. *Syntactic complexity:* Syntactic complexity encompasses the range and sophistication of syntactic features, typically measured in terms of unit lengths, amount of subordination/coordination, range, and degree of sophistication (Lu, 2011). Since syntactic complexity is not the focus of the research, 'large-grained indices of syntactic complexity' were selected which have a positive relationship with language proficiency (Kyle & Crossley, 2017), specifically two measures which can differentiate between proficiency levels (Juffs et al., 2020; Lu, 2010, 2011; Wolfe-Quintero et al., 1998). These measures

were calculated using the Web-based L2 Syntactic Complexity Analyzer software (Lu, 2010).

a. *Complex nominals per clause (CN/C)*: a complex nominal is an expression consisting of a head noun with a preceding noun or adjective.

b. *Mean length of clause (MLC)*: a clause is defined as a group of words containing a subject and a predicate, and length is measured by the number of words.

In addition, two grammatical accuracy metrics were calculated – grammatical errors per 100 words and punctuation errors per 100 words, as manually identified by two linguists (myself and a professor).

6. *Collocational proficiency*: operationalized using the three Collgram indices (Bestgen & Granger, 2014; Granger & Bestgen, 2014) described in Section 3.3.2 and two additional metrics to capture collocational accuracy. All five measures were calculated in Python using frequency information from COCA. The collocation span is 4 (4L and 4R):

   a. *mean MI*: to measure association of collocations containing infrequent words using the formula from Davies (2008-):

   $MI = \log\left((W1, W2\ freq * Corpus\ size) / (W1\ freq * W2\ freq * span) \right) / \log(2)$

   b. *mean t-score*: to measure association of collocations containing high frequency words using the formula from Evert (2009): $\text{t} - \text{score} = (O - E)/\sqrt{O}$ where O is the observed frequency and E is the expected frequency, calculated as $E = (W1\ freq * W2\ freq) / Corpus\ size$.

   c. *absent bigrams*: the proportion of bigrams absent from the reference corpus calculated as *Number of absent bigrams in text / Total bigrams in text*.

   d. *Number of accurate collocations and collocation errors (per 100 words)*: Only lexical word choice errors were considered so that, for example, "*put food <u>in the table</u>*" (B2,

para. 3) is considered a collocational error with the preposition *in*, but "*they had <u>everything give</u>*" (B1, para. 2) is not tagged despite the incorrect form of the verb *give*. This policy of selecting lexical collocational inaccuracies aligns with previous studies focusing on acceptable multi-word units (e.g., Crossley et al., 2015) and the lexical/collocational error types identified by Granger (2003) and Wanner et al. (2013). Collocations taken from the task prompt, e.g. *deal with (the) problems*, were not counted.

7. *Collocation frequency bands*: it was determined for each collocation whether it contained only high-frequency words, a mid-frequency word, or a low-frequency word, i.e., a collocate frequency approach was employed (see Section 3.3.1). The proportion of each of these types of collocations in the text was then calculated.

Importantly, collocations were identified using a combined phraseological and frequency-based approach (see Section 3.2.1). First, collocations were manually tagged independently, using a checklist of phraseological criteria adapted from previous studies (Appendix E). These criteria relate to the collocation length (2-3 words), the types of words (at least one lexical word), and their nature as a single 'chunk'. Frequency criteria were then used to settle any disagreements and to check any word strings which the annotators flagged as potential collocations. The COCA frequency criteria used were raw frequency > 5 and an association score of MI > 3 or t-score > 2.

5.2.2 Normalization findings

For each of the indices, the normalized texts were within 5% of the original texts (Table 1). Of the indices selected, the ones which correspond most clearly to proficiency are AG

(sophistication), CN/C and MLC (syntactic complexity), number of grammar errors and collocation errors (accuracy), and mean MI and number of correct collocations (collocation use). As such, we might expect these indices to have the greatest impact on human ratings in the subsequent study.

Table 1 Comparison of original and normalized texts

text	length	Lexical indices			Syntactic indices			
		vocD	*AG*	*bigram range*	*CN/C*	*MLC*	*gram errors*	*punc errors*
B1 orig	172	48.28	0.38	0.09	0.62	6.62	5.81	7.56
B1 norm	250	48.66	0.38	0.09	0.64	6.41	6.10	7.60
B2 orig	349	46.26	0.90	0.16	0.86	7.29	2.87	2.58
B2 norm	250	44.17	0.94	0.17	0.80	7.31	2.80	2.40
C1 orig	254	70.23	1.44	0.12	2.05	11.77	0.79	3.15
C1 norm	250	70.51	1.45	0.12	2.05	11.59	0.80	3.20

text	Collocational indices							
	mean MI	*absent bigrams*	*mean t-score*	*low freq*	*mid freq*	*high freq*	*col errors*	*correct cols*
B1orig	2.43	0.98	115.84	0.0	0.25	0.75	8.14	4.66
B1 norm	2.52	0.99	111.67	0.0	0.25	0.75	8.00	4.80
B2 orig	2.94	0.96	162.09	0.0	0.26	0.74	3.15	8.88
B2 norm	2.88	0.96	153.30	0.0	0.27	0.73	3.20	8.80
C1 orig	3.26	0.93	93.62	0.09	0.36	0.55	1.97	12.99
C1 norm	3.23	0.93	92.03	0.09	0.36	0.55	2.00	13.20

In addition to adjusting for length, spelling was also normalized by removing orthographic errors. Removal of spelling errors is a common step in cleaning texts prior to automated analyses (e.g., Bestgen & Granger, 2014; Vögelin et al., 2019) since spelling significantly affects accuracy scores (Polio & Shea, 2014). In fact, there are only two spelling errors in the original texts, both in B1: *arounds (around) and childrens (children)*. These errors may also be due to a lack of morphological awareness, but as they both result in a non-word, here they are labeled as spelling errors. In either case, as the B1 text is at the bottom of the B1 range (IELTS Band 4), correcting these two errors would certainly not impact the overall ratings to any great extent.

5.3 Base texts

In creating the normalized texts, the number of accurate and inaccurate collocations per 100 words was controlled for, i.e., the collocational density. However, for experimental purposes, the distributions of these collocations were not ideal. For example, in the normalized B1, B2, and C1 texts, there are 20, 8, and 5 inaccurate collocations, with the B2 and C1 texts much more alike in this regard. Therefore, the normalized texts were minimally adjusted to create 'base' texts with more evenly spaced collocational use. Table 2 presents a comparison of the text versions, with the base texts highlighted. Here we see that the accurate collocations essentially remain unchanged, with only one accurate C1 collocation altered to be inaccurate. As a result, the accurate collocations per 100 words increases by 4.0 at each CEFR level (B1 = 4.8, B2 = 8.8, C1 = 12.8). For the inaccurate collocations, two were removed from B1 and four were added to B2 to lessen the difference between the B1 and B2 collocation profiles. Based on these changes, the number of inaccurate collocations per 100 words decreases by 2.4 at each level (B1 = 7.2, B2 = 4.8, C1 = 2.4).

Table 2 Collocational density of text versions

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

5.4 Expert ratings

The true test of comparability between the texts is whether or not human ratings remain consistent. As mentioned, the ratings for the original texts are provided by IELTS. The ratings for the normalized and base texts come from IELTS examiners (Figure 4 and Figure 5). The characteristics of these raters and these ratings will be discussed in detail in Chapter 6, but for now we will take as a given their validity based on the fact that (a) the raters are assessment experts, (b) each text was rated multiple times, and (c) ratings were calibrated using MFRM models.

The overall scores for each text are presented in Figure 4. Overall scores were calculated as an average of the four analytic bands of Task Response (TR), Coherence and Cohesion (CC), Lexical Resource (LR), and Grammatical Range and Accuracy (GRA). Following IELTS assessment practices, scores were rounded down to the nearest 0.5 so that 6.25 → 6.0 and 6.75 → 6.5. These scores indicate that at all three proficiency levels, the normalization process did not significantly impact the average ratings. For B1, the scores increased from 4.0 to 4.4, for B2 they decreased from 6.5 to 6.0, and for C1 they decreased from 8.0 to 7.5. In other words, the changes

are all within half a band. The same is true of the base texts. At B1 and C1, the base texts received the same scores as the original texts, and at B2 the score is 0.5 lower, matching the normalized text score. Importantly, the overall CEFR level was maintained for all three levels and across all three text versions (see Figure 4).

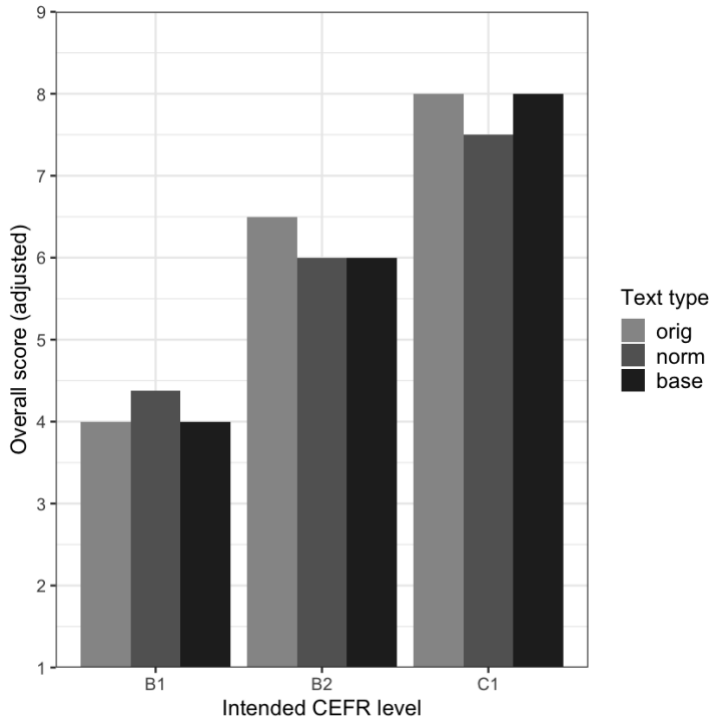

Figure 4 Overall ratings comparison of original, normalized, and base texts

A closer look at the analytic scores (Figure 5) indicates that the overall scores are truly comparable; the overall scores are not fortunate averages from jagged profiles, but match the trends seen in the individual analytic scores. In the majority of cases, analytic scores are identical for the three text versions, and when there is a discrepancy, it is an increase of one band in a B1 original text score (TR and LR) or a decrease of one band in a B2/C1 original text score (TR for B2, CC and GRA for C1). For analytic scores, a band difference of one is the minimum difference possible

as half points cannot be awarded. There are also no evident patterns of one band descriptor being less reliable than the other, with either one or two discrepancies for each category.

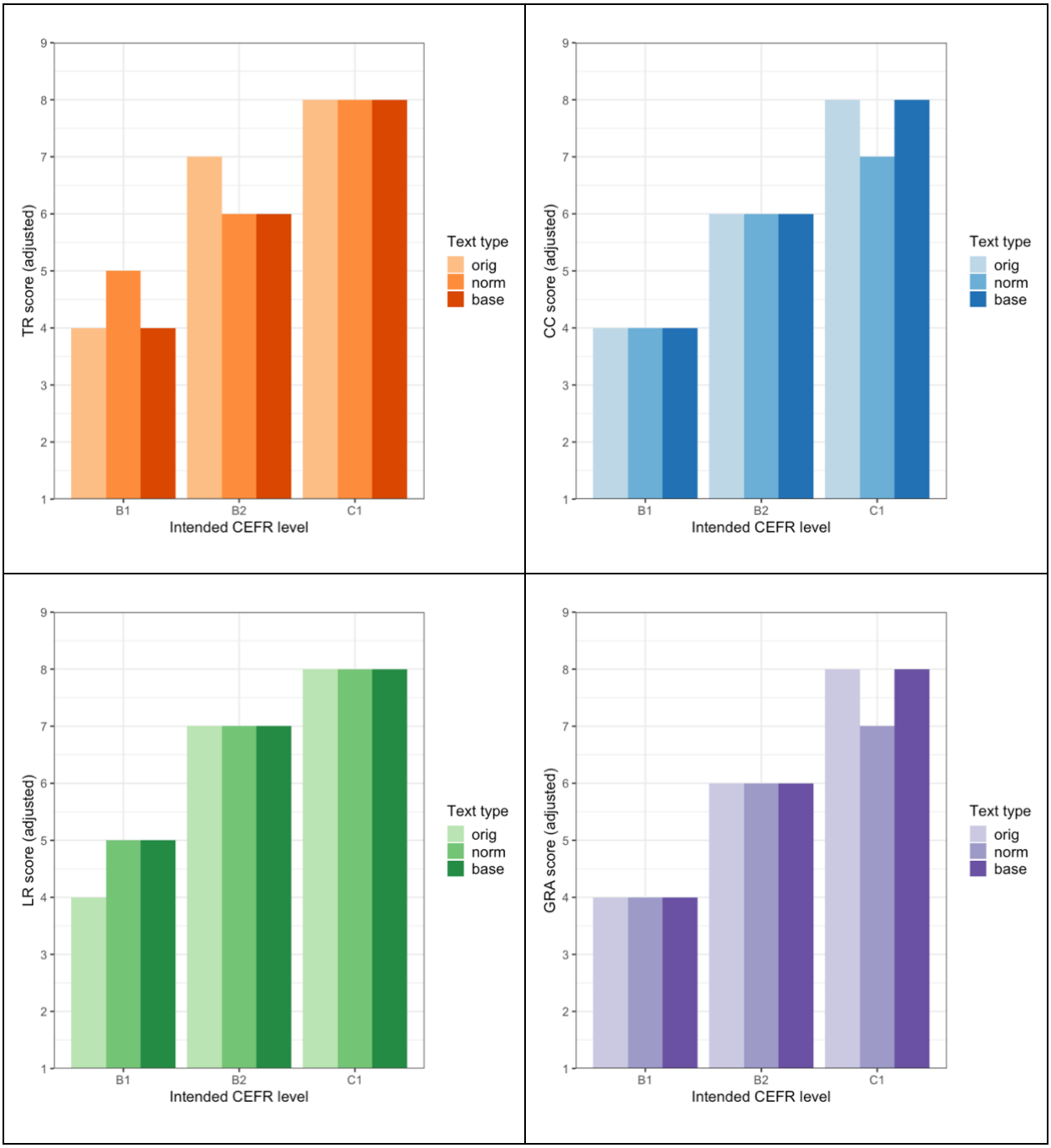

Figure 5 Analytic ratings comparison of original, normalized, and base texts

5.5 Discussion and chapter summary

In this chapter, we have looked in detail at the characteristics of three texts which will form the basis of the upcoming research in Chapter 6. By normalizing the length of these three texts, the aim was to eliminate an important variable to ensure maximum validity (Mickan, 2003). In preparing the normalized texts, numerous lexical and grammatical statistical measures were monitored to ensure minimum differences between the original and normalized versions. Subsequently, the base texts were developed to create a more even distribution of accurate and inaccurate collocations across CEFR levels. The ratings by assessment experts validate these changes because there were minimal differences in the overall and analytic band scores. As a result, we can conclude that the normalized and base texts represent the same CEFR proficiency levels as the original texts.

6.0 Study: Expert ratings of collocational proficiency

This chapter presents the design, materials, and findings of an experiment that is based on the considerations introduced in Chapters 2-4 and uses the three base texts presented in Chapter 5. In brief, from the base texts, a set of 30 versions were produced, manipulating specific variables relating to collocational sophistication and collocational accuracy. In the second stage, 47 IELTS examiners rated the texts holistically and analytically. They also answered questions designed to elicit their perceptions of salient lexical features that led to their assessments. The goals of the experiment were to isolate and measure the contributions of collocational features to overall ratings of essay quality by comparing quantitative text metrics and expert ratings. Specifically, three research questions were addressed:

1. To what extent do statistical measures of collocational proficiency predict expert rater judgments of overall and lexical proficiency in terms of:
    a. collocational sophistication?
    b. collocational accuracy?
2. Is the distinction between, 'low', 'mid', and 'high' frequency lexical items as part of collocations significant?
3. To what extent do expert raters' scores align with the features of the texts that they stated as being most salient/impactful?

6.1 Methodology

In the study, statistical text analysis is followed up with human ratings and reflections. The research therefore can be said to integrate quantitative and qualitative linguistic data within a single study (Bryman, 2006; Creswell & Plano Clark, 2011; Hashemi, 2012), using an embedded design in which both types of data are collected simultaneously (Creswell & Plano Clark, 2011). While it is tempting to therefore claim that the study adopts a mixed-methods approach, this is not the case; as Davis (1995, p. 435) explains, "the use of qualitative techniques does not constitute the approach." Rather, the quantitative component (the ratings) provides the primary data, and the qualitative technique (the questionnaire) enhances the 'completeness' of the data by adding further explanation of the rating process (Bryman, 2006). In the field of assessment research, quantitative studies are the norm, but incorporating a qualitative dimension is valuable for understanding raters' decision-making processes (Goh & Ang-Aw, 2018).

6.1.1 Instruments

All of the instruments were presented to raters remotely through a Qualtrics survey (Qualtrics, 2020). To start, raters were provided a link for viewing/downloading the rating scales and task prompt. Next, the raters scored three texts. Once a rater submitted a score, they could not go back and later change it. After rating each text, the raters answered follow-up questions about their assessment decisions. Finally, raters provided personal metadata.

6.1.1.1 Text modifications

Using the three base texts from Chapter 5, 30 different versions were created (10 versions per base text; Appendix F) by changing up to approximately 12% of the words (30/250). These manipulations were intended to influence the collocational indices introduced in Section 5.2.1, resulting in four different variables: proficiency level, collocational sophistication, non-collocational sophistication, and collocational accuracy. What follows are descriptions of these variables and examples of the changes made:

1. *Proficiency level:* There are three CEFR proficiency levels, B1 (intermediate), B2 (upper-intermediate), C1 (advanced). See Chapter 5 for details.
2. *Collocational sophistication:* There are three levels, 'Low', 'Mid', and 'High' sophistication. To change the levels of sophistication, accurate collocations were replaced based on the lemma frequencies of the collocates (see Section 3.3.1). To do so, the frequency bands of collocates in COCA were calculated, and these were then verified using the learner corpus frequency list PSL3 (Naismith et al., 2018).
    a. *Low:* all lemmas in the collocation are in the K1-2 frequency bands in each version, e.g.,
        - (B1) 'good effect'
        - (B2) 'support the idea'
        - (C1) 'learn to survive'

        It was not necessary to make changes to create additional K1-2 collocations.
    b. *Mid*: at least one lemma in the collocation is in the K3-9 frequency bands, and there are no K10+ lemmas, e.g.,
        - (B1) 'in my case' → 'for me personally'
          K1 K1 K1 K1 K1 K3

- (B2) ‘work very hard’ → ‘work incredibly hard’
  K1 K1 K1 K1 K4 K1
- (C1) ‘good example’ → ‘concrete example’
  K1 K1 K4 K1

c. *High:* at least one lemma in the collocation is in the K10-16 frequency bands, e.g.,

- (B1) ‘hear about’ → ‘see firsthand’
  K1 K1 K1 K11
- (B2) ‘nice clothes’ → ‘trendy clothes’
  K1 K1 K11 K1
- (C1) ‘feel cheated’ → ‘feel victimized’
  K1 K4 K1 K13

3. *Non-collocational sophistication:* The same characteristics of collocational sophistication apply to non-collocational sophistication. The only difference is that the words altered are not part of collocations.

a. *Low:* Again, it was not necessary to make changes to create additional K1-2 lemmas. Examples of such lemmas include ‘however’ (B1), ‘something’ (B2)’, and ‘problem’ (C1).

b. *Mid*: change of non-collocation lemmas to K3-9:

- (B1) ‘strong’ → ‘mature’
  K1 K6
- (B2) ‘think’ → ‘presume’
  K1 K6
- (C1) ‘see’ → ‘witness’
  K1 K4

c. *High:* change of non-collocation lemmas to K10-16:

- (B1) ‘nevertheless’ → ‘unbelievably’
  K4 K14
- (B2) ‘toys’ → ‘belongings’
  K3 K10
- (C1) ‘talent’ → ‘ingenuity’
  K2 K13

4. *Collocational accuracy:* There are two accuracy levels, 'High' and 'Low'. At each proficiency level, there are six additional inaccurate collocations in the low level. As a result, the number of inaccurate collocations in 'High' B1 is equal to the number of inaccurate collocations in 'Low' B2 and so on, providing an even distribution of collocation errors (Table 3).

Table 3 Accurate and inaccurate collocations in 'High' and 'Low' accuracy texts

Text	Total collocations	Accurate collocations	Inaccurate collocations
B1 Low accuracy	30 (100%)	12 (40%)	18 (60%)
B1 High accuracy	30 (100%)	18 (60%)	12 (40%)
B2 Low accuracy	34 (100%)	22 (65%)	12 (35%)
B2 High accuracy	34 (100%)	28 (82%)	6 (18%)
C1 Low accuracy	38 (100%)	32 (84%)	6 (16%)
C1 High accuracy	38 (100%)	38 (100%)	0 (0%)

Table 4 presents a matrix of all 30 text versions. As in Lenko-Szymanska (2019, p. 161), the overarching selection criteria for the indices was the "meaningfulness and interpretability of the information they encapsulate as well as their theoretical motivation", i.e., they correlate with human judgements of proficiency (see Chapter 4). The indices also align with the IELTS Task 2 band descriptors (Appendix G; IELTS, n.d.-c, Writing task 2 assessment criteria). For example, the Lexical Resource band descriptors refer to vocabulary range at every band, including 'a limited range' (B5), 'an adequate range' (B6), 'a sufficient range' (B7) and 'a wide range' (B8) of vocabulary. Range in this case is one component of sophistication, i.e., diversity. The other component, the 'advanced' nature of the vocabulary is also represented as 'basic vocabulary' (B4), 'less common vocabulary' (B6), 'less common lexical items' (B7), and 'uncommon lexical items'

(B8), though exact frequency bands are not specified. Description of accuracy is also included at every level, including in reference to word choice, collocation, and word formation. From the complete set of 30 text versions, each examiner rated three texts, one randomly assigned from each proficiency level to avoid rating multiple texts derived from the same base.

Table 4 Characteristics of text versions

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

6.1.1.2 Rating scales

Recall from Section 4.1.3. that a combination of holistic and analytic scales is more appropriate than holistic scales alone, allowing for more systematic and detailed assessment. To rate the texts analytically, raters therefore used the IELTS public writing scales (Appendix G; IELTS, n.d.-c, Writing task 2 assessment criteria). Importantly, these public scales closely match the confidential IELTS scales in terms of categories and descriptors. For each of the four categories – Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy – there are bands from 1-9 with positive, descriptive criteria. To achieve a Band score, a writer must fully achieve all of the points in that band. For this study, the 9-point band scale was further divided into three sub-levels, e.g., 5-, 5, 5+, so that there was a 'strong' and 'weak' possibility within each band (a practice used in Jarvis, 2013b). This alteration was intended to prevent 'bunching' whereby there is little variance in the ratings from version to version (Wallace, 2009), a trend found in Lumley (1998).[9] Should bunching occur, the data would not reveal actual differences in raters' opinions based on the lexical features. In addition, the reliability of scales with more items is typically higher (DeVellis, 2003), and it has been suggested that researchers start with twice as many scale levels as will eventually be used (Morgado et al., 2017). Of particular interest for this research is the score of the Lexical Resource category, and the overall score (the average of the four categories).

In addition to the analytic scales, the raters also provided a holistic assessment (Appendix H) based on the IELTS public 9-band overall scale (IELTS, n.d.-b). IELTS examiners do not give holistic assessments, but this additional holistic rating serves to align the methods of the current study with other comparable research and provides an extra level of data for analysis. The holistic

[9] Worldwide, the mean academic writing score is 5.6 (IELTS, 2018).

scales were minimally adapted to remove reference to spoken production and language comprehension.

After submitting a rating for a text, raters were asked the following question to elicit a rationale for their Lexical Range score and to reveal which lexical elements were most salient/consciously noticed:

> *What features did you notice which led to your rating of Lexical Resource?*
>
> *Please comment on any general features or specific lexical items which impacted your decision.*

6.1.2 Raters

For the purposes of this study, more important than the *exam* itself is the characteristics of the *examiners*; the research is intended to pertain not only to characteristics of IELTS Writing tasks, but rather, to English L2 academic writing in general and how it is perceived by assessment experts. As was observed in the literature review, the expertise of raters plays an important factor in determining ratings outcomes. To this end, the raters in this study are all current or former IELTS writing examiners, selected due to the process by which all IELTS examiners are recruited and trained.

To be eligible for training, prospective examiners must have substantial (typically 3+ years) teaching experience to adults, an undergraduate degree, and a recognized TEFL/TESOL qualification or degree in education. As Lumley (1998) notes, using experienced language teachers as raters is a common practice. Applicants must also possess expert spoken and written proficiency in English (Band 9 on IELTS). The application/interview process is followed by an induction and training process which lasts two days (for the Writing section only), at the end of which

certification takes place through independent assessment of a set of test texts. Once active, examiners are subsequently monitored at least three times in their first year through double marking and feedback on their examining. Failure to achieve the required standards results in further monitoring, training, and ultimately loss of examiner status. Every two years, a further one day of standardization and re-certification is required to maintain examiner status. Taken as a whole, the above quality control process can be seen to adhere to many of the best practices described in Section 4.1 for making writing assessment as objective and reliable as possible.

To recruit the examiners for this study, the Snowball Sampling Method (SSM) method was used, also known as chain-referral sampling. Essentially, SSM is a type of sampling of convenience (Cohen & Arieli, 2011) in which the researcher recruits one participant who in turn contacts others in their social network, and so on, so that the sample increasingly 'snowballs' in size. SSM may be considered less optimal compared to random sampling methods because the samples derived from SSM are more likely to be biased or unrepresentative (Browne, 2005). However, in certain circumstances, SSM is the most practical choice, especially when recruiting members from hard-to-reach groups (Valdez & Kaplan, 1998).

IELTS examiners fit this 'hard-to-reach' description as it is a small population, spread thinly throughout the world, with no formal registry which can be publicly accessed. Furthermore, as suggested by Cohen and Arieli (2011), parallel snowball networks were initiated, i.e., multiple snowballs starting with unrelated raters, to increase the sampling representativity. In SSM there are a variety of ways to initiate first contact (Browne, 2005), and in this case potential participants were contacted directly via email and social media. This method was possible due to my own professional experience: I was an active IELTS examiner, senior examiner, online examiner, and examiner trainer for 14 years. Importantly, in these roles I worked extensively at different centers

in four countries (online and face-to-face), resulting in personal connections to examiners and IELTS administrators around the world. As such, I can been seen as 'embedded' in the social network (Browne, 2005).

Ultimately, the purpose of using this methodology was to collect as much data as possible from a relatively small population. This goal was achieved as 47 suitable examiners responded, well beyond the suggested rule-of-thumb minimum of 30 proposed for reliable statistical reasons (Brown, 1988). One participant was excluded as they appeared to be a non-examiner who mistakenly believed they were eligible. This anonymous participant wrote 'other' in the examiner status field and selected 'less than one year experience' (there was no 'zero experience' option). Each examiner assessed three texts and answered follow-up questions (approximately 20 minutes total). This participant pool size had the desired effect of allowing multiple raters for each script, which increases reliability and objectivity (Robitzsch & Steinfeld, 2018).

Table 5 presents the raters' demographic information. These data present a picture of a group of mature examiners in terms of age (all over 30) and experience, both in terms of TESOL experience (94% have over 10 years' experience) and examining experience (85% have more than 3+ years' experience). Of the 47 participants, 19% are examiner trainers, held to a higher standard of reliability. Most commonly, the participants' experience has been with students with a wide range of first languages (60%) and proficiency levels (70%). The participants are also highly trained, with most possessing graduate degrees (85%) and additional TESOL certification (89%). The majority of the examiners are former, rather than current examiners, which is unsurprising given a number of factors: many exam centers have closed due to the COVID-19 pandemic, some centers do not allow active examiners to participate in such studies, and many centers now send

the writing tests to central hubs to be marked rather than by local examiners. However, their lapsed status does not invalidate their previous extensive assessment experience and qualifications.

Table 5 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	Senior examiner and trainer			
	38	7	2			
Current status	Active	Lapsed				
	9	38				
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	Wide	Narrow				
	28	19				

6.2 Findings

6.2.1 MFRM model

Recall from Section 4.3 that the purpose of using MFRM models for this study was to arrive at fair scores for each of the criteria for each of the texts. These fair scores can then be used in subsequent statistical analyses to reliably represent the ratings of the average, unbiased rater.

For all of the MFRM analysis, FACETS software was used. To start, a 3-facet (i.e., 3-variable) model was created consisting of the text, the rater, and the criterion. The data were transformed into a 'long format' such that each rating was considered a unique datapoint. In total, there were 705 datapoints produced (47 raters x 3 texts x 5 criteria). There are a number of assumptions for MFRM models and accompanying statistics to check the properties of the rating scales (McNamara et al., 2019). In this study, the IELTS rating scales themselves are not under investigation and therefore their validity will not be examined using this small dataset when extensive validation studies have been conducted by the examining body itself. Instead, data were examined to identify raters whose scores were outliers.

However, of importance here for the MFRM results is the assumption that there should be more than ten observations per category. In these data, each of the 27 possible ratings (9 bands with three sub-bands, e.g., -1, 1, 1+ … 9-, 9, 9+) met this assumption except for the lowest ratings awarded with ratings of 3 and 3+ receiving five observations each. This sparsity of ratings in band 3 indicates that they are being awarded so infrequently as to essentially change the 9-band scale into a 6-band scale, which was to be expected given the known CEFR levels of the three base texts. This conclusion is supported by the number rating strata (5.28) indicated in the Rater Measurement Report model (Table 6).

It is also necessary to examine the rater facet in terms of reliability, separation, standard error (SE) and fit (Aryadoust & Liu, 2015). Reliability refers to the reproducibility of the estimates if a different sample from the same population was rated. The reliability statistic in Table 6 is between 0 and 1, with higher coefficients signifying higher reproducibility. The summary of these data indicates high reliability of 0.94. Separation measures the number of distinct strata of severity (Erguvan & Aksu Dunya, 2020), in this case there are 3.90 for the sample, with individual raters ranging between -0.43 to +0.50 log-odd units (SE = 0.05) in terms of severity. This finding means that there are distinct groups of raters in terms of severity, which can negatively or positively impact ratings by up to approximately half a band. There are also two 'fit' statistics which indicate the difference between the observed and expected responses, i.e., the degree to which the data accurately fit the model. *Infit* is calculated based on average item difficulty and *outfit* on high and low item difficulty. 'Items' in this case refers to assessment criteria, and the very little difference between the two statistics indicates that certain criteria were not rated more severely than others. For these fit statistics, scores closer to 1 are more reliable, with a reasonable range between 0.5 and 1.5; scores under 0.5 indicate less variation than expected whereas scores over 1.5 indicate a greater degree of divergence from expectations and are more problematic (Aryadoust & Liu, 2015; Linacre, 2020).

The significance of these differences is confirmed by the chi-square statistic which tests whether the severity levels between raters is significant (Erguvan & Aksu Dunya, 2020). Taken together these results add further evidence that that the ratings should be treated prior to further analysis. To address raters/ratings that did not fit, outliers were removed. In FACETS, outliers are reported as 'unexpected responses' and are defined as responses whose standardized residuals are outside ±3. In these data, 22 datapoints were highlighted (3.1%) and omitted before re-running the

MFRM analysis. There was no pattern for the outliers in terms of which category was being rated, but two raters in particular were the source of many of the outlying data points, though after removing these outliers no further action needed to be taken.

Table 6 Summary measurement results for the rater facet

***n* = 47**	**Prof. measure**	**SE**	**Infit**	**Outfit**	**Fair avg.**	**Obs. avg.**	**# of ratings**
Mean	0.00	0.05	0.98	0.99	64.20	65.08	15
S. D. (population)	0.19	0.00	0.54	0.53	6.15	5.39	0
S. D. (sample)	0.19	0.00	0.55	0.53	6.21	5.45	0
Model (population)	SD. = 0.18, Separation = 3.86, Strata = 5.48, Reliability = 0.94						
Model (sample)	SD. = 0.19, Separation = 3.90, Strata = 5.54, Reliability = 0.94						
Model, Fixed (all same)	χ^2 = 717.6, df = 46, p = 0.00						
Inter-rater agreement	opportunities = 1175, Exact agreements = 137 (11.7%), Expected = 129.3 (11.0%)						

Having analyzed and cleaned the data, the next step was to obtain the fair scores. To do so, the procedure from the FACETS manual was followed (Linacre, 2020). Essentially, the values of all facets except for the ratings are set to their mean values and the analysis is performed. In completing this process, the expected average for each rating is then the same as the fair score (Linacre, 2020), and the equation from Section 4.3 becomes the following (Eckes, 2009, p. 22):

$$\ln\left[\frac{p_{nk}}{p_{nk-1}}\right] = \theta_n - \beta_M - \alpha_M - \tau_k$$

Table 7 presents an overview of the scores for each text. Here the differences between the overall fair average scores and the observed average scores can be seen (the average scores are the sum of the scores for each of the five categories), as well as the other infit/outfit statistics previously discussed. The individual fair scores for each of the five criteria are also provided. For example, the first line represents the scores for Text 1 (B1, low sophistication low Accuracy). With a proficiency measure of -0.64 logits (*SE* = 0.5), this text received near the lowest ratings. The Infit/Outfit measures of 0.56/0.55 indicate that these data had satisfactory model fit, both with and without outliers as they fall between 0.5 and 1.5. The 'Observed average' of 4.97 is the mean rating

for this text across for all 20 ratings (see '# of ratings' column), i.e., four examiners rated this text, giving one rating for each of the five analytic bands. The 'Fair average' is slightly lower at 4.92, which means that the raters of Text 1 were slightly lenient. More usefully, the individual criteria fair scores are then provided, ranging between 4.8 and 5.3; it is these scores which will be used in subsequent analyses. As one would expect, the 'Proficiency measure', 'Observed average', 'Fair average', and fair scores generally rise across the 30 text versions, especially when comparing the B1 texts (Texts 1-10), the B2 texts (Texts 11-20) and the C1 texts (Texts 21-30).

Table 7 Measurement results for the text facet

Text	Prof. measure	SE	Infit	Outfit	Fair avg.	Obs. avg.	# of ratings	TR fair	CC fair	LR fair	GRA fair	HOL fair
1	-0.64	0.05	0.56	0.55	49.22	49.70	20	4.9	4.8	5.3	4.8	4.8
2	-0.64	0.05	1.07	1.09	49.11	47.11	19	4.9	4.8	5.2	4.8	4.8
3	-0.70	0.05	1.76	1.75	47.89	48.26	19	4.8	4.6	5.1	4.6	4.7
4	-0.63	0.05	3.05	2.95	49.34	45.45	20	4.9	4.8	5.3	4.8	4.8
5	-0.76	0.05	1.22	1.23	46.49	48.85	20	4.7	4.5	5.0	4.5	4.5
6	-0.51	0.04	0.83	0.86	51.66	53.38	24	5.2	5.0	5.5	5.0	5.1
7	-0.54	0.05	0.92	0.90	51.15	51.56	18	5.1	5.0	5.4	5.0	5.0
8	-0.45	0.04	0.93	0.92	52.92	50.80	25	5.3	5.2	5.7	5.2	5.2
9	-0.58	0.04	0.68	0.68	50.44	47.44	25	5.1	4.9	5.4	4.9	4.9
10	-0.54	0.05	0.91	0.90	51.13	53.20	20	5.1	5.0	5.4	5.0	5.0
11	0.10	0.04	0.74	0.75	66.43	67.55	20	6.7	6.5	7.1	6.5	6.5
12	-0.05	0.04	0.82	0.83	62.59	63.15	20	6.3	6.1	6.7	6.1	6.1
13	0.04	0.04	0.66	0.70	64.91	64.89	19	6.5	6.3	7.0	6.3	6.4
14	0.12	0.04	0.74	0.72	67.12	67.30	20	6.7	6.5	7.2	6.5	6.6
15	0.10	0.04	1.01	1.02	66.45	69.15	20	6.7	6.5	7.1	6.5	6.5
16	0.27	0.04	0.79	0.78	71.29	66.70	20	7.1	6.9	7.6	7.0	7.0
17	0.01	0.04	0.52	0.52	63.96	61.07	30	6.4	6.2	6.9	6.2	6.3
18	0.00	0.04	0.60	0.56	63.72	61.85	20	6.4	6.2	6.9	6.2	6.2
19	0.20	0.04	1.19	1.20	69.28	71.65	20	6.9	6.7	7.4	6.7	6.8
20	-0.18	0.04	0.52	0.52	59.14	63.82	22	5.9	5.7	6.4	5.7	5.8
21	0.76	0.04	0.73	0.72	85.33	82.05	19	8.6	8.4	8.9	8.4	8.4
22	0.53	0.04	0.52	0.49	78.84	83.80	20	7.9	7.7	8.4	7.7	7.7
23	0.63	0.04	1.43	1.44	81.74	80.75	20	8.2	8.0	8.6	8.0	8.0
24	0.69	0.04	1.33	1.34	83.40	82.15	20	8.4	8.2	8.8	8.2	8.2
25	0.43	0.04	0.87	0.88	75.84	78.55	20	7.6	7.4	8.1	7.4	7.4
26	0.81	0.05	0.84	0.78	86.45	84.68	22	8.7	8.5	9.0	8.5	8.5
27	0.68	0.04	1.12	1.11	83.08	77.95	20	8.3	8.1	8.8	8.1	8.2
28	0.29	0.04	1.34	1.31	71.84	72.39	18	7.2	7.0	7.7	7.0	7.0
29	0.43	0.04	0.69	0.70	75.96	79.25	24	7.6	7.4	8.1	7.4	7.4
30	0.79	0.05	1.29	1.30	85.99	87.00	20	8.6	8.4	9.0	8.4	8.5

In addition to the three main facets, demographic information about the raters was collected: the 11 variables in Table 5 as well as time spent on the rating task and order of texts rated. It was therefore possible to examine whether these 13 variables affected the ratings using FACETS' *Bias analysis* in which a secondary analysis is performed for each specified bias interaction. These bias analyses do not affect the calculated fair scores but can help understand why groups of raters may be performing differently. For these data, one additional dummy facet was added at a time and the bias analysis report considered. Variables with less than 0.5 in the probability column indicate significant bias (McNamara et al., 2019), and with this threshold, none of the 13 investigated variables were significant.

6.2.2 Correlation analysis

Having established fair scores for each text for each criterion, the impact of collocational features on the lexical and holistic scores can be calculated (30 texts, 128 sets of ratings). Starting with a more holistic overview, the dependent variable (Lexical Resource) and the experimental variables of sophistication, sophistication type, accuracy, and proficiency level (CEFR) were checked for possible correlations using a correlation matrix (Figure 6). In addition, variables predicted to be important to the model were included, namely, the other three analytic scores (Task Response, Coherence and Cohesion, and Grammatical Range and Accuracy). The results of this initial correlation analysis are not illuminating: There is a very strong correlation between all of the analytic and holistic scores, increasing with the initial CEFR level. In this figure, no significant correlations are apparent with respect to the experimental variables (e.g., between sophistication type and high sophistication, $r = 0.02$, $p = 0.83$), indicating a need for more fine-grained analysis using models in which potential effects are not masked by other factors or interactions.

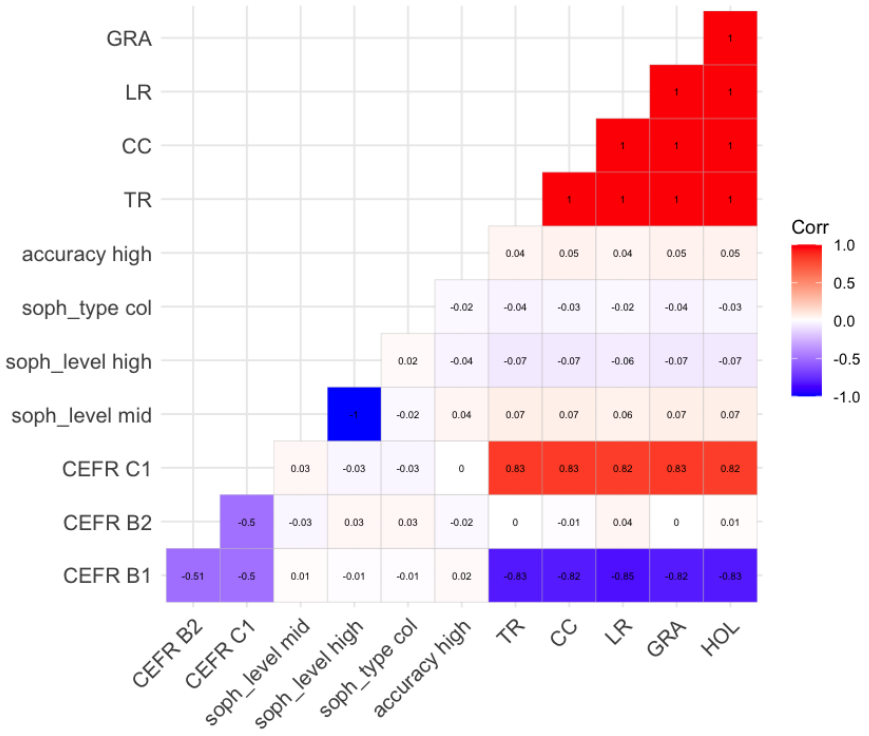

Figure 6 Correlation matrix with variables of interest

6.2.3 Linear regression models (LRMs)

Further inferential analysis was performed through three linear regression models (LRMs). The LRMs were created in the R environment (version 3.6.2; R Development Core Team, 2019) using the lm package. Often, linear mixed effect models are appropriate for ratings data to account for random variation amongst participants. Here, the MFRM fair scores negate this need, and there were no other significant demographic or task factors which might have been potential random effects.

Prior to creating the models, the assumptions required by linear regressions were checked (Levshina, 2015): the observations are independent as there is no predicted clustering of data points beyond the factors to be included in the model. Within these data points, the outcome (or response)

variables of the models are considered to be interval-scaled (see Section 4.1.3). Homoscedasticity of variance and the relationship between variables was checked by plotting the residuals against fitted values. For normality of residuals, Q-Q plots based on z-scores were created. To check for collinearity a statistical test showed that Corrected Variance Inflation Factor was less than five for sophistication and accuracy; as expected there was collinearity between the different analytic bands, but this is not an issue as these factors are not under investigation (Allison, 2012; O'Brien, 2017). Likewise, a low Durbin-Watson statistic (0.58) indicated autocorrelation, which was expected given the relationship between CEFR levels, holistic scores, and analytic band scores; however, this assumption is not critical for studies that do not use time-sensitive data (as in this case). Based on these observations and tests, it was confirmed that all necessary assumptions were met.

6.2.3.1 Lexical Resource LRM (experimental design)

In the first model (Table 8), the outcome variable is the Lexical Resource fair scores (LR_fair), i.e., how the lexical resource score varied or was predicted by other variables. These independent variables are the fair scores for the other analytic criteria (TR, CC, GRA), the level of sophistication (Low, Mid, High), the type of sophistication (Collocation, Non-collocation), the collocation accuracy (Low, High), and the base text CEFR level (B1, B2, C1). In addition, motivated interactions are included: CEFR:sophistication, CEFR:accuracy, sophistication:accuracy, and sophistication:sophistication type. In this first experimental design, all potential variables are included in the model regardless of whether they improve the model fit. All of the independent variables were sum contrast coded (as in Picoral & Carvalho, 2020) though the sophistication contrasts were eventually dropped because there is no sophistication type associated with low sophistication texts. The reference level for sophistication is therefore ‘Low’.

As a result, the model's intercept is dispersed across levels of the other variables. These data partially answer research question 1 that asks the extent to which statistical measures of collocational sophistication and accuracy predict ratings of lexical proficiency.

Table 8 Linear regression model for factors predicting Lexical Resource ratings (experimental design)

Parameters	*Estimate*	*SE*	*CI*	*t value*	*Pr(>|t|)*
(Intercept)	0.831	0.069	0.69 – 0.97	11.975	**<0.001*** **
CEFR [B1-C1]	-0.180	0.018	-0.22 – -0.15	-10.174	**<0.001*** **
CEFR [B2-C1]	0.070	0.007	0.06 – 0.08	10.479	**<0.001*** **
TR fair	0.883	0.144	0.60 – 1.17	6.118	**<0.001*** **
CC fair	-0.827	0.262	-1.35 – -0.31	-3.151	**0.002** **
GRA fair	0.879	0.199	0.48 – 1.27	4.422	**<0.001*** **
soph [high]	0.032	0.006	0.02 – 0.04	5.104	**<0.001*** **
accuracy [low-high]	-0.005	0.005	-0.01 – 0.00	-1.057	0.293
soph type [col-non-col]	-0.022	0.004	-0.03 – -0.01	-4.887	**<0.001****
CEFR [B1-C1] * soph [high]	-0.014	0.010	-0.03 – 0.01	-1.411	0.162
CEFR [B2-C1] * soph [high]	-0.010	0.009	-0.03 – 0.01	-1.182	0.240
CEFR [B1-C1] * accuracy [low-high]	-0.030	0.005	-0.04 – -0.02	-6.454	**<0.001*** **
CEFR [B2-C1] * accuracy [low-high]	0.001	0.005	-0.01 – 0.01	0.320	0.750
soph [high] * accuracy [low-high]	0.009	0.006	-0.00 – 0.02	1.452	0.150
soph [high] * soph type [col-non-col]	0.008	0.006	-0.00 – 0.02	1.367	0.175

Note. * $p<.05$, ** $p<.01$, *** $p<.001$
Model formula: lm(formula = LR_fair ~ CEFR + TR_fair + CC_fair + GRA_fair + soph + accuracy + soph_type + CEFR:soph + CEFR:accuracy + soph:accuracy + soph:soph_type)
$R^2 = 0.99$

Looking at the parameters in Table 8, we see that CEFR level (rows 2 and 3) is significant. However, because of the combination of contrast and treatment coding, these estimates are difficult to interpret. Instead, the post-hoc analysis in Table 9 (Tukey's multiple comparison test) confirms that the levels are reliably different, increasing as expected from B1 → B2 → C1.

Table 9 Tukey's multiple comparison of means test for CEFR (LR experimental design)

Contrast	*Estimate*	*SE*	*CI*	*df*	*t ratio*	*Pr(>|t|)*
B1 - B2	0.252	0.017	0.22 - 0.29	89	-14.551	**<0.001*** **
B1 - C1	0.310	0.031	0.25 - 0.37	89	-9.952	**<0.001*** **
B2 - C1	0.058	0.017	0.00 - 0.06	89	-3.398	**<0.003** **

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

Next, we see that the TR, CC, and GRA criteria also all significantly impact LR, the potential reasons for which will be discussed further in Section 6.3.4. Generally, it can be said that higher ratings in other analytic bands positively affect LR ratings, whether because of a halo effect, overlapping characteristics, or other potential interactions which might affect the ability of examiners to separate these constructs for the purposes of assessment.

Considering sophistication, there is a significant positive increase, i.e., higher sophistication predicts a higher LR rating. To better understand the sophistication results, a Tukey's multiple comparison post-hoc test was again run (Table 10). Here we see that the bulk of the sophistication effect occurs when going from low sophistication to high sophistication. The difference between low and mid sophistication is also significant (p = 0.018), but there is no significant difference between mid and high sophistication (p = 0.158).

Table 10 Tukey's multiple comparison of means test for sophistication (LR experimental design)

Contrast	*Estimate*	*SE*	*CI*	*df*	*t ratio*	*Pr(>|t|)*
mid vs. low	0.032	0.012	0.01 - 0.06	111	2.810	**0.018***
high vs. low	0.050	0.010	0.03 -0.07	111	4.831	**<0.001*** **
high vs. mid	0.018	0.009	0.00 - 0.04	111	1.934	0.158

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

In addition, there is a significant difference for sophistication type (Table 11); whether sophistication increased within or outside of collocations was meaningful in these data. This effect is small but significant, resulting in higher Lexical Resource scores when the sophisticated words were part of a collocation. It must be remembered that with the context of learners' writing, there may be no truly 'individual words', with every word a part of at least one construction (and consisting of morpheme constructions themselves). However, that sophisticated words within collocations specifically are significant indicates that collocations are especially salient to examiners as a construction type. Together, these findings answer the second research question which asked whether there was a significant distinction between, low-, mid-, and high-frequency lexical items.

Table 11 Tukey's multiple comparison of means test for sophistication type (LR experimental design)

Contrast	*Estimate*	*SE*	*CI*	*df*	*t ratio*	*Pr(>|t|)*
col vs. non-col	0.036	0.006	0.024 - 0.05	89	5.744	**0.018***

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

Ultimately, of the experimental variables, only collocational accuracy was not significant. It was predicted that the 'High' accuracy level would lead to higher LR ratings, but this was not the case. There was one significant interaction between the low-high accuracy contrast and the CEFR B1-C1 contrast. However, after careful plotting and examination of this interaction, this significant relationship appears to be spurious.

6.2.3.2 Lexical Resource LRM (exploratory design)

A second *exploratory* LRM with LR as the outcome variable was also created to investigate what the best-fitting model would be involving the variables from the experimental model

described above. In this case, the model was created using stepwise model selection, as recommended in Baayen (2008) and Monteiro et al. (2020), wherein the initial model formula was the outcome variable (LR fair score) and one independent variable (CEFR level). One by one, independent variables and interactions were added to see if they were significant ($p<.05$ for the individual coefficients within the model) and improved the model fit (likelihood ratio test). The resulting best-fitting model was nearly identical to the experimental model and had the same R^2 but it did not include the variable of accuracy, the interaction of sophistication:accuracy, or the interaction of sophistication:sophistication type. In other words, this exploratory model confirmed that the eliminated variables were not reliable predictors in the experimental model.

6.2.3.3 Holistic rating LRM (experimental design)

The third LRM (Table 12) is identical to the first experimental model except that the outcome variable is the holistic fair score (HOL_fair) instead of Lexical Resource, which then becomes an independent variable. The purpose of this model is to better answer the first research question, this time investigating the effect of collocational sophistication, sophistication type, and accuracy on more global holistic ratings. Overall, many of the same patterns are present as in the Lexical Resource LRM, albeit with smaller effects, as one might expect, because Lexical Resource is but one component of the holistic score. In terms of the other criteria, because HOL takes into consideration the overall text quality, it would be logical for all four analytic criteria to positively predict the HOL rating. However, only Task Response (TR) was significant, indicating that the manner in which the writer addresses the prompt outweighs other linguistic features in determining the holistic impression. In this model, sophistication is again significant and accuracy is again non-significant. In contrast to the earlier model, sophistication type only approaches significance (0.08), suggesting that although sophistication type, that is whether inside or outside a collocation,

meaningfully impacts lexical scores, the effect size is minimal when judging overall text quality. From the post-hoc test (Table 13) the same trends are present that were seen with LR in Table 10. Here, however, the significance for the difference in means between Mid and Low sophistication only borders on significance ($p = 0.051$).

Table 12 Linear regression model for factors predicting Holistic ratings (experimental design)

Parameters	*Estimate*	*SE*	*CI*	*t value*	*Pr(>|t|)*
(Intercept)	-0.479	0.136	-0.75 – -0.21	-3.531	**0.001*** **
CEFR [B1-C1]	0.123	0.032	0.06 – 0.19	3.900	**<0.001*** **
CEFR [B2-C1]	0.007	0.012	-0.02 – 0.03	0.586	0.559
TR fair	0.545	0.208	0.13 – 0.96	2.617	**0.010***
CC fair	0.098	0.335	-0.57 – 0.76	0.293	0.770
LR fair	0.120	0.128	-0.14 – 0.37	0.934	0.353
GRA fair	0.290	0.266	-0.24 – 0.82	1.093	0.278
soph [high]	0.015	0.009	-0.00 – 0.03	1.783	0.078.
accuracy [low-high]	-0.001	0.006	-0.01 – 0.01	-0.097	0.923
soph type [col-non-col]	-0.015	0.006	-0.03 – -0.00	-2.548	**0.013***
CEFR [B1-C1] * soph [high]	-0.052	0.012	-0.08 – -0.03	-4.305	**<0.001*** **
CEFR [B2-C1] * soph [high]	0.013	0.011	-0.01 – 0.03	1.227	0.223
CEFR [B1-C1] * accuracy [low-high]	-0.007	0.007	-0.02 – 0.01	-1.054	0.295
CEFR [B2-C1] * accuracy [low-high]	0.004	0.006	-0.01 – 0.01	0.642	0.522
soph [high] * accuracy [low-high]	-0.002	0.007	-0.02 – 0.01	-0.214	0.831
soph [high] * soph type [col-non-col]	0.017	0.007	0.00 – 0.03	2.492	**0.015***

Note. * $p<.05$, ** $p<.01$, *** $p<.001$
Model formula: lm(formula = HOL_fair ~ CEFR + TR_fair + CC_fair + LR_fair + GRA_fair + soph + accuracy + soph_type + CEFR:soph + CEFR:accuracy + soph:accuracy + soph:soph_type)
$R^2 = 0.99$

Table 13 Tukey's multiple comparison of means test for sophistication levels (HOL experimental design)

Contrast	*Estimate*	*SE*	*CI*	*df*	*t ratio*	*Pr(>|t|)*
mid vs. low	0.022	0.009	0.00 - 0.04	110	2.417	0.051.
high vs. low	0.032	0.009	0.01 - 0.05	110	3.673	**0.001****
high vs. mid	0.010	0.007	0.00 - 0.01	110	1.402	0.415

Note. . $p<.1$, * $p<.05$, ** $p<.01$, *** $p<.001$

6.2.4 Rater comments

As previously described, there is one supplemental qualitative data component to the ratings: the raters' rationales for their scores. These data are intended to answer the third research question which asked to what extent expert raters' scores align with the features of the texts that they stated as being most salient/impactful.

Overall, the comments contained both positive and negative elements, contrasting Hall and Sheyholislami (2013) who recorded three times more negative comments than positive. There was a great deal of variety in terms of the level of detail of comments, ranging from 'good vocabulary' (Text 1) or 'very good LR' (Text 2) to much longer responses with lists of specific examples. Raters also routinely used language directly from the band descriptors, e.g., "uses a limited range of vocabulary which is minimally adequate" (Text 7) or "uses a wide range of vocabulary fluently and flexibly to convey precise meanings" (Text 19).

In terms of what the raters consciously noticed, Figure 7 presents a tally of the different lexical features commented on. In these counts, similar terms were collapsed into the following categories:

- *appropriacy*: appropriacy, adequacy, flexibility, precision, relevance
- *error gravity*: impact on communication, strain on reader, difficulty in understanding

- *range*: range, repetition, variety
- *sophistication*: sophistication, how common/rare the lexis is
- *style:* register, style, tone
- *formulaic sequences*: chunks, idioms, phrasal verbs, noun phrase
- *coherence and cohesion*: coherence, cohesion, linking words

When accuracy was mentioned, if the type of accuracy was specified, a unique tag was created, resulting in tags for (general) accuracy, word choice accuracy, and collocational accuracy. One other element noted six times was the B1 and B2 writers' reliance on lexis from the rubric. This was not included in Figure 7 as it is not a specific lexical feature, but it could be justifiably included in the 'range' category.

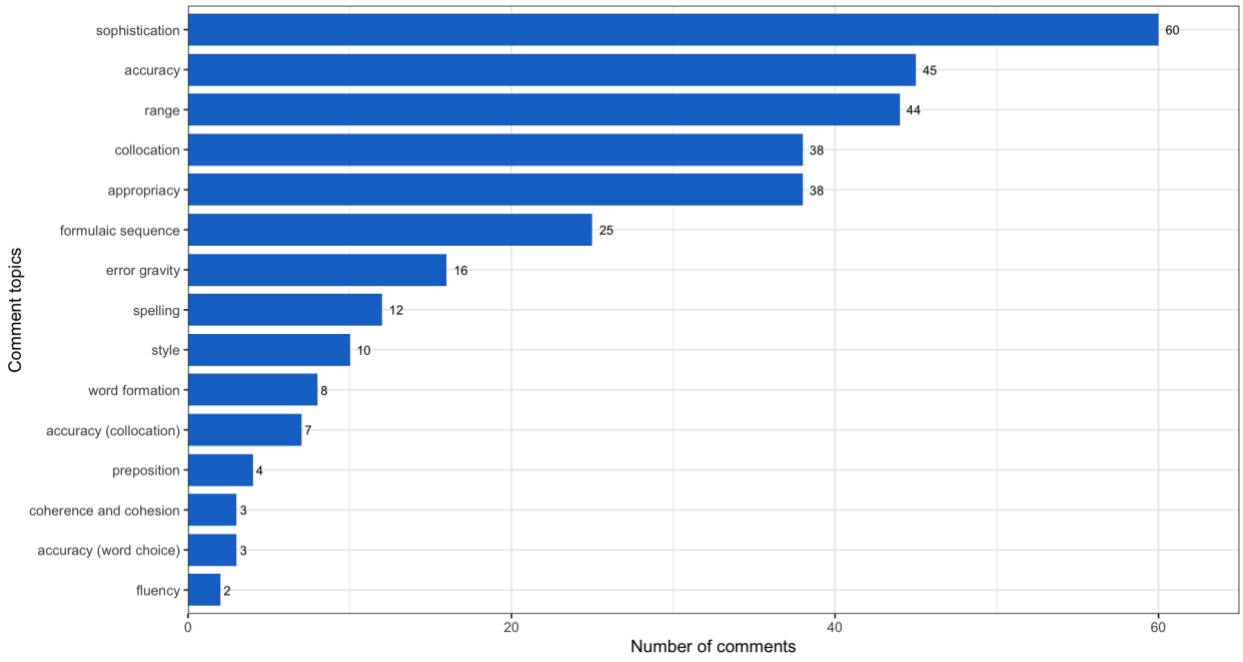

Figure 7 Topics of rater comments

As seen in Figure 7, 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. What is more, when giving examples, raters tended to give a mix of single and multi-word sequences, suggesting that collocations were salient even if the term 'collocation' was not explicitly used in their comments. In the examples, both individual words and collocations were provided, though more collocations than single words. In some cases, a specific collocate appeared to be particularly salient as some examiners gave the single word as an example and others gave the word as part of a collocation, e.g. 'errand' vs. 'do errands' or 'first-hand' vs. 'first-hand experience'. In the same vein, prepositions appeared to be important to examiners both as individual words and as part of collocations. These cases exemplify how raters may differ in the extent to which multiword items are noticed compared to single words.

The examples given also provide support for the statistical importance of sophistication. Across all three levels, low-frequency single words and collocations containing low-frequency words were flagged as being examples of sophisticated lexis. For example, sophisticated single words repeatedly highlighted include 'tremendous' (Text 3), 'fantasize' (Text 10), and 'flaunts' (Text 20). Sophisticated collocations included 'fairly young' (Text 4), 'sense of respect' (Text 16), and 'sheer motivation' (Text 25). These examples suggest that the experts' lexical awareness of frequency is high in terms of single words or collocations. This finding might support the importance of considering the collocate frequency approach (Section 3.3.1) since it does not seem that low-frequency collocations with high-frequency collocates were noticed by the raters other than one idiomatic phrase, 'put food on the table' (Text 14). That many of the manipulated items were noticed as being collocations also supports the research's experimental design and the psycholinguistic validity of these collocations as *bona fide* chunks. Likewise, many raters said that lexis was the strongest element of the writing for the B1 and B2 texts, suggesting that the

sophistication manipulations noticeably affected the LR scores to a greater extent than the other criteria.

Considering accuracy, we may recall from the LRMs that collocational accuracy was not statistically significant in predicting LR or HOL scores. However, the raters' comments demonstrate that lexical accuracy in its many forms is a feature they consider important. For one, many raters commented on the high accuracy of the spelling. In terms of collocational accuracy, collocations with inaccurate word choice were frequently noted, e.g. 'positive school' (Text 3), 'study at money' (Text 6), 'impact to a child' (Text 16), and straight contribution (Text 25). As a result, the writers were often described as 'risk takers' (Texts 3, 9, 14, 15, 19, 27), i.e., writers with higher sophistication but lower accuracy. This characterization occurred at all three CEFR levels:

- (B1) "There were a few flashes of less common language… but these were far outweighed by the frequent errors causing strain for the reader"
- (B2) "very specific and high level usage... not all accurate, so this is why it's not a band 9."
- (C1) "They are aware of less common lexis and idiomatic phrases… but accuracy of use prevents a higher grade."

6.3 Unexpected findings

6.3.1 Lack of accuracy significance

As described in this chapter, collocational accuracy was not significant in predicting lexical or holistic ratings. And yet, in their comments, the raters' commonly described accuracy, including collocational accuracy, as being one of the determining factors for their decisions. What is more,

there is a consensus of studies into lexical accuracy (Section 2.3) and collocational accuracy (Section 3.3.3) that both types of accuracy significantly impact text quality, and it remains a key element of assessment descriptors for tests like IELTS. It therefore seems unlikely that collocational accuracy is not an important element for models of lexical proficiency.

There are two potential explanations for the discrepancy between the expected and actual results with regards to accuracy. The first is that the quantity of collocation errors between the 'Low accuracy' and 'High accuracy' versions was insufficient to be impactful. 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 (Low/High accuracy: B1 = 18/12, B2 = 12/6, C1 = 6/0). For example, the Band 6 descriptor says "uses less common lexical items with some awareness of style and collocation"; it is feasible that both 18 and 12 collocation errors are subsumed under '*some* awareness'.

A second explanation relates to the level of *error gravity* (Section 2.3), i.e., the impact of the errors on communication. In this case, both the original and inserted collocation errors can be considered to have low error gravity as at all three proficiency levels the meaning of the inaccurate collocations is still clear, e.g.,

- (B1) *positive school* → *excellent school*
- (B2) *set their mind that* → *decide for themselves that*
- (C1) *in the weekends* → *on the weekend*

In this study, error gravity was not statistically measured, e.g., through a judgement test, but it was intentionally controlled for by substituting collocates to maintain similar meanings across versions. Thus, error gravity was unlikely to be a confounding variable, but the consistent 'light' error gravity likely decreased the impact of the collocation inaccuracies.

Both of the above explanations are reasonable when we consider the training that L2 language teachers (who then become examiners) receive with respect to error correction. Communicative Language Teaching (CLT) in its many incarnations has been the dominant paradigm in English language teaching since the 1970s (Thornbury, 2016). Central to CLT is the notion of *communicative competence* (Hymes, 1972), i.e., the ability to communicate appropriately in different social contexts. By emphasizing communicative goals, the quantity and gravity of errors are given less weight as long as they do not impact communication.

6.3.2 Relationship between low-, mid-, and high- sophistication

It was originally hypothesized that there would be a clear distinction between low sophistication (use of K1-K2 bands), mid sophistication (K3-9 bands), and high sophistication (K10-16 bands). The results somewhat confirmed these initial expectations since the variable of sophistication did predict higher lexical ratings. However, the difference between mid and high sophistication was not significant ($p = 0.158$, $t = 1.934$).

Recall from the discussion of lexical sophistication (Section 2.2.2) that advanced words are often defined as being outside the K2 or K3 bands. Superficially, these results could be seen to support this view since low-mid and low-high contrasts were significant but mid-high was not. However, if we consider that the greatest contrast is between low-high and that even the mid-high is meaningful (though not significant at $p<0.05$), then a second interpretation of these results is that sophistication in terms of frequency is better seen on a cline. Yes, more sophisticated lexis is less frequent, as confirmed by the overall sophistication variable and the post-hoc comparisons. However, the division of 'basic' and 'advanced' words which is incorporated into sophistication

measures like AG may not capture the increasing sophistication of lexis across frequency bands, which themselves are quite wide with 1000 lemmas each.

6.3.3 Relationship between Lexical Resource and other analytic ratings

Task Response (TR), Coherence and Cohesion (CC), and Grammar (GRA) were significant variables in predicting Lexical Resource (LR) ratings. One possibility is the presence of a halo effect, with stronger ratings in one category unintentionally biasing ratings in another. For TR, this relationship is logical as TR is not a distinct linguistic system; effective use of relevant lexis directly affects a writers' ability to convey the intended message as required by a specific task prompt. However, the relationship between LR/CC and LR/GRA is less clear-cut.

With respect to CC, it may be that this category impacts LR due to the use of discourse markers; these phrases overlap both categories because on the one hand they affect coherence and cohesion and on the other hand they are salient multiword phrases which often contain less frequent words. Conklin and Schmitt (2012) suggest that signalling discourse organization is in fact one of the key communicative functions of FSs. Plus, L2 English teachers are trained to value discourse markers as indicators of good writing. As a result, the use of less-frequent discourse markers such as 'for me personally' (replacing 'in my case') may be responsible for increases in ratings for both criteria.

As for GRA, the separation between grammar and lexis has always been a contentious issue (see Section 3.2.3). On the one hand, these findings could be used as evidence that lexical and grammatical assessment categories should be merged (Römer, 2017; Ruegg et al., 2011) , in line with the construction grammar stance that syntax and lexis are merely descriptions of different construction types with different levels of complexity and abstraction (Goldberg, 2013; Halliday

& Matthiessen, 2014). Alternatively, proponents of maintaining the grammar/lexis distinction (e.g., Bulté & Housen, 2014; Foster & Tavakoli, 2009; Skehan, 2009) could justifiably claim that the class of collocations are unique in that they do contain grammatical and lexical elements. As such, how to assess collocations perhaps needs greater attention in training and assessment materials, without completely abandoning the grammar/lexis dichotomy which has high perceived validity and is a mainstay of L2 English curricula, materials, and training.

6.4 Chapter summary

This chapter has described the methodology of the primary study which involved collecting ratings and comments from a pool of expert examiners. The findings were then analyzed using MFRM and Linear Regression models and contextualized with the raters' descriptive comments. These triangulated data depict a complex relationship between different aspects of lexical and collocational proficiency. Notably, it appears clear that lexical sophistication in general is impactful, especially when sophisticated words are part of salient collocations. However, the exact divisions between 'low', 'mid' and 'high' are not clear cut, especially between 'mid' and 'high'. Furthermore, collocational accuracy seems to be noticeable to raters, but did not impact the statistical models.

7.0 Study: Discussion and conclusions

This final chapter considers the research's implications in terms of two main areas: language teaching and language assessment. The chapter then concludes by summarizing the research, highlighting some of the limitations, and suggesting fruitful avenues for future investigations. Overall, this work can be seen to mirror the goals of Vögelin et al. (2019, p. 58) in that "this research is embedded in a larger effort to identify key factors influencing assessment processes to improve classroom assessment and teacher education."

7.1 Pedagogical implications

The importance of collocation instruction spans all levels, modes, and types of English. After all, collocations are ubiquitous no matter the context (Section 3.1). As well, recall from Section 3.2.3 that from a Construction Grammar perspective, collocations are one type of construction, but can also contain or be part of other constructions. Thus, by giving pedagogical attention to collocations, learners can strengthen form-meaning mappings for morphemes, words, collocations, and idioms. What is more, this type of lexical/collocational knowledge also facilitates grammatical expansion of longer constructions which initially include such collocations (Zyzik & Gass, 2008).

Considering how to best develop academic English writing proficiency is particularly important due to the challenges of acquiring expertise in this domain and the number of learners striving for this goal. For example, there were close to one million international students in the US

in the 2020-2021 academic year (Open Doors, 2021). In such settings, students require a threshold proficiency to avoid academic failure which would negatively impact both the students and the schools (Roche & Harrington, 2014).

7.1.1 How to teach collocations

It is beyond the scope of this paper to consider the best specific methods for teaching collocations. However, from the large body of research in this area, a few general consensuses are pertinent to the current discussion. First, it is not sufficient to leave collocation learning up to chance exposure, especially if opportunities for exposure are limited (Granger, 1998). In general, learners do not receive sufficient input to learn more sophisticated vocabulary incidentally (Cobb & Laufer, 2021); for example, Nation (2014) estimates that it would require a learner to read one million words in order to increase their receptive vocabulary knowledge from 4,000 to 5,000 word families, a figure which is unrealistic for the timeframes of most learners.

Rather, explicit vocabulary instruction is necessary, i.e., any specific strategy that increases learners' depth of lexical knowledge (Lee, 2003). Numerous authors have reiterated the importance of explicit vocabulary and collocation instruction (e.g., Dabbagh & Janebi Enayat, 2019; Fan, 2009; Koda, 1993; Pellicer-Sánchez, 2020; Rossiter et al., 2016), with research showing that such instruction leads to improved receptive and productive knowledge (e.g., Boers et al., 2017; Lee, 2003; Webb & Kagimoto, 2011). Explicit collocation instruction can take many forms, including teacher-led clarification or guided discover to promote noticing, but it also encompasses a wide variety of task types from controlled gap-fills, to freer narrative production tasks, to translation activities. Regardless of the specific task types, it is necessary that realistic goals be set and the importance of collocations be emphasized (Levitzky-Aviad & Laufer, 2013).

That is not to say that incidental input is not also important; lexical items must be encountered on multiple occasions, ideally in a range of contexts, for them to be acquired (Nation, 2020; Webb et al., 2013). Even then, however, collocations are especially slow to be incidentally learned as learners may focus more on the meaning of individual content words than multiword sequences (Boers, 2020). A variety of options exist to accelerate incidental collocation learning including *flooding* (increasing the number of occurrences in a text), *enhanced output* (making the target language more salient, e.g. through highlighting), and *chunking* (marking the boundaries of formulaic sequences). The results of these interventions are somewhat mixed, with flooding more notably improving receptive than productive knowledge (Pellicer-Sánchez, 2017; Webb et al., 2013), enhanced output increasing students' noticing of the target language (Boers et al., 2017) but decreasing attention on text contents (Choi, 2017), and chunking only producing a minor impact (Lewis, 1993; Nation, 2011). Overall, a combination of incidental and explicit learning is a pragmatic route to acquiring collocational proficiency. As Macis and Schmitt (2016, p. 15) summarize in relation to learning collocational meanings, "[the research] seems to suggest good old-fashioned educational values: going to school and reading a lot."

7.1.2 Choosing which collocations to teach

If we are in agreement that explicit collocation learning is essential for English learners, the question becomes which collocations to teach, and it is this endeavor which the findings from the previous chapter can most help inform. Since the publication of *The Lexical Approach* (Lewis, 1993), there has been "a growing lexicalization of the teaching syllabus" (Bestgen & Granger, 2014, p. 29), and at this point collocation instruction is common in many contexts and materials. However, the selection of which collocations to teach often remains unprincipled (Macis &

Schmitt, 2016). A general rule-of-thumb is to consider the cost-benefit principle so that learners get the best return for their effort during their limited time for studying whether in class or at home (Coxhead, 2000, p. 213; Laufer & Nation, 2012).

As we have seen, frequency is one way of deciding this benefit and has been traditionally used with reference to coverage (Section 2.2.3). Students start by learning the most frequent vocabulary and then move to less frequent vocabulary, thereby increasing the number of words they can likely comprehend in a text. For students studying English for Academic Purposes or in academic contexts, mid-frequency lexis is considered to be particularly important to learn (Nation & Anthony, 2013; Schmitt, 2010; Vilkaitė-Lozdienė & Schmitt, 2020) as it is essential for achieving sufficient coverage of academic texts (Laufer, 1989; Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006)

To facilitate this process, General English word lists provide a useful shortcut (Coxhead, 2020; Nation & Waring, 1997). For example, lists of individual words include the popular New General Service List (NGSL; Browne et al., 2013) or the more recent Nuclear Word Family List (NFL7; Cobb & Laufer, 2021). A very limited number of collocation frequency lists have also been produced, e.g., Shin and Nation (2007) compiled a short list of the 100 most frequent collocations in British spoken English. More commonly, there are longer collocation dictionaries for learners, e.g. the Oxford Collocation Dictionary (McIntosh et al., 2009). The lack of collocation lists is understandable; it is hard to make one which is comprehensive and practical (Wolter, 2020) given the number of collocations which exist and the relatively low frequency of collocations compared to individual words. There have been calls for frequency lists which incorporate both individual words and formulaic sequences, but to date no such list has been created (Vilkaitė-Lozdienė & Schmitt, 2020). Another downside to both word and collocation lists is that they

typically only cover the most frequent items, i.e., low sophistication, with everything else considered ‘advanced’ by process of elimination.

Another popular approach to selecting high-benefit lexis is by using the principles of frequency and list making to focus exclusively on academic words and collocations. Academic vocabulary can be broadly defined as lexis which is common in academic discourse but infrequent in General English (Charles & Pecorari, 2016), accounting from 10-14% of lexis in academic written texts (Coxhead, 2000; Gardner & Davies, 2014). Such vocabulary is therefore spread out over different frequency bands in a general reference corpus, including K10+ items (Coxhead, 2020). The Academic Word List (AWL; Coxhead, 2000) is a testament to the usefulness of this approach as it has been widely adopted in EAP contexts. As the AWL relates to collocations, Coxhead (2020) essentially advocates for a collocate frequency approach by suggesting a focus on complex noun phrases containing AWL words or AWL words which co-occur such as ‘analysis and assessment’.

Taking a collocation frequency approach to EAP list creation, Ackermann and Chen (2013) developed the Academic Collocation List (ACL). This list is comprised of 2468 academic collocations which were selected based on their collocation frequency in the Pearson International Corpus of Academic English (PICAE) and were subsequently evaluated on phraseological and pedagogical criteria by a panel of experts. Similarly, Durrant (2009) produced a 1000-word EAP collocation list based on academic vocabulary in a bespoke corpus of research articles. Interestingly, Durrant found minimal overlap between the words in his collocation list and the words in the AWL. This divergence points to the different results that are obtained depending on whether one adopts a collocate frequency approach (as in Cox, 2020) or a collocation frequency approach (as in Durrant, 2009).

Thus, we have seen that words and collocations may be selected using a haphazard approach or a frequency/list approach (whether general or from a specialized domain like Academic English). Considering the results of the present study, we must consider that the impact of frequency on productive lexical proficiency and receptive ability is different. In terms of academic writing, as sophistication increases, so too does lexical and overall text quality, with the greatest difference between low and high sophistication. These findings would therefore suggest that the inclusion of some lexis in the K10+ bands in learning goals is 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."

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 three categories represent high sophistication collocations which nonetheless have fairly wide academic generalizability and therefore high cost-benefit:

1. Discourse markers:
   - 'in contrast' → 'in startling contrast' (K10)
   - 'in my case' (K1) → 'not to generalize' (K13)
   - 'together' (K1) → 'in conjunction' (K16)

Recall, the Lexical Resource LRM indicated the interplay between the categories of 'Lexical Resource' and other analytic criteria such as 'Coherence and Cohesion' and 'Grammatical

Range and Accuracy). By replacing or inserting K10+ words into discourse markers, students can apply these FSs in a range of academic text types to good effect, concurrently improving the sophistication of their lexis, the demonstrating flexible use of cohesive devices, and in some cases increasing grammatical complexity.

2. Synonyms for other K1-2 collocations:

- ‘very important’ → ‘of paramount importance’
  K1 K1 K11 K2
- ‘learn about’ → ‘gain proficiency in’
  K1 K1 K1 K10 K1
- ‘come together’ → ‘result in a convergence’
  K1 K1 K2 K1 K1 K11

Not only do these examples include K10+ words, but they also promote nominalization. Collocations containing nouns are the most frequent type of lexical collocation (Nizonkiza & Van de Poel, 2019) and are a key attribute of academic prose (Biber & Gray, 2013b; Wells, 1960). However, learners tend to underuse noun forms in their own writing in favor of verbs (Naismith & Juffs, 2021; Parent, 2019), indicating that more explicit instruction in this area is warranted.

3. Personally relevant, domain-specific, specialized lexis:

- ‘*be sick*’ → ‘*suffer from an ailment*’ (medicine)
  K1 K1 K2 K1 K1 K11
- ‘*smart shopper*’ → ‘*savvy shopper*’ (marketing)
  K1 K11
- ‘*vary their portfolio*’ → ‘*diversify their portfolio*’ (finance)
  K3 K11

Many students enroll in EAP programs with the intention of progressing to an academic degree program instructed in English. For such students, it is necessary to not only know general academic English vocabulary, but also lexis specific to their desired future studies and careers.

This type of *specialized vocabulary* is necessary for success in English medium programs (e.g., Coxhead, 2020; Nation, 2013), and it is also one of the greatest challenges that learners report (Dang & Dang, 2021).

There may of course be a number of other categories of collocations which are pedagogically worthwhile at an individual, class, or institutional level, and which might be explored in future research. What is important is that findings from lexical research of this type be made accessible and palatable to teachers so that they can be applied in the classroom. To this end, there are practical text analysis tools which are freely available online and which can be used to investigate lexical items to assess their suitability.

One option is to start with a text which is to be used in class, for example from a coursebook or an authentic source. This text can then be inputted into a web-based tool to determine the frequency/sophistication level of each word. Depending on the teaching context, different tools may be of more use. For example, Lextutor can be used to see the K-band of each word, the Online Graded Text Editor (OGTE; Waring & Browne, n.d.) to see which words occur within a specific vocabulary list, and Text Inspector (EnglishProfile, 2015) to see the CEFR level of each word. As an example, the introduction from the Wikipedia entry for 'Pittsburgh' ("Pittsburgh", 2021) was analyzed through each of the three websites. Table 14 presents the words which OGTE marked as being beyond 'Mid-near-native (~8000 headwords)'.

Table 14 Comparison of word classifications from online text tools

OGTE words	Lextutor classification	Text Inspector classification
populous	K10	Off-list
confluence	K10	Off-list
skyscraper	Off-list	Off-list
automotive	K8	Off-list
stockholder	Off-list	Off-list
(per) capita	K12	Off-list
transform	K3	B2

Here we see that there is considerable overlap in the evaluations of this lexis with Lextutor agreeing that the majority of the words are high sophistication (‘populous’, ‘confluence’, ‘per capita’), approaching high sophistication (‘automotive’), or off-list and cannot be assigned a K-band. Likewise, for all but one word, Text Inspector assigned the ‘off-list’ label, in this case indicating that the words are rare and infrequently used by even C2 level learners. One word, ‘transform’ was labeled as rare by OGTE but not by Lextutor or Text Inspector. Equipped with this type of frequency information, a teacher can then select items which align with the needs of their learners, and using a phraseological approach, can identify the collocations within which these items are situated. In this case, from the text the collocations ‘populous city’, ‘confluence of’, and ‘automotive industry’ might be appropriate targets for instruction and useful for the learners’ future academic writing. Of note, the online tools considered here primarily provide frequency information for individual words and thus align with the collocate frequency approach. At present there are no known common online tools for recognizing all collocations in a text and their collocational frequencies.

A second option is to take a bottom-up approach by starting with the desired lexis and looking up collocations to find ones at a suitable sophistication level. These collocations can then either be taught directly or included in input, e.g. when designing course materials. For example, imagine a scenario teaching EAP to students who are interested in pursuing a degree in health sciences. Using the online COCA interface, adjective collocations with ‘disease’ can be queried, as in Figure 8 (1-2 words left of the node, sorted by MI). From this list, good high-sophistication candidates for instruction include ‘transmitted’ (K15), ‘degenerative’ (K21), and ‘autoimmune’ (K15) which may be generalizable to a range of medical contexts or ailments.

HELP			FREQ		ALL	%	MI
1	☐	CARDIOVASCULAR	1780		5802	30.68	10.74
2	☐	CHRONIC	1579		18285	8.64	8.92
3	☐	INFECTIOUS	1330		6388	20.82	10.19
4	☐	CORONARY	1255		3315	37.86	11.05
5	☐	CELIAC	540		1142	47.29	11.37
6	☐	MAD	467		39890	1.17	6.03
7	☐	AUTOIMMUNE	452		1780	25.39	10.47
8	☐	RARE	429		42512	1.01	5.82
9	☐	INFLAMMATORY	413		4454	9.27	9.02
10	☐	PULMONARY	405		3068	13.20	9.53
11	☐	OBSTRUCTIVE	348		966	36.02	10.98
12	☐	TRANSMITTED	341		6544	5.21	8.19
13	☐	FATAL	331		11537	2.87	7.33
14	☐	DEADLY	321		17584	1.83	6.67
15	☐	DEGENERATIVE	306		976	31.35	10.78

Figure 8 Adjective collocations with 'disease' in COCA

In summary, the availability of free online tools means that teachers have ready access to a wealth of lexical frequency information. Whether starting with the text or with the lexis, teachers can therefore make choices regarding which lexical items to teach, combining frequency information with their own expertise, intuitions, and knowledge of their students' needs.

7.2 Assessment implications

The lexical aspects analyzed in this study can also contribute to the discussion of how lexical proficiency should be assessed, specifically in terms of rating scales and rater training.

7.2.1 Rating scale implications

With respect to the public IELTS Lexical Resource descriptors, we see that relativistic terminology is frequently used to distinguish between bands. For example, sophistication in bands

6 to 8 includes the descriptions "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 too. Research has shown that teachers debate the meanings of terms in descriptors (Claire, 2001) and have difficulty interpreting/applying relativistic terminology (Smith, 2000). However, more prescriptive criteria lead to less agreement (Smith, 2000). A compromise could therefore perhaps 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 concrete examples.

Next, we consider how collocations (and formulaic sequences more generally) are represented in these analytic scales. Table 15 presents how dimensions of lexical proficiency are addressed at each band level.

Table 15 Lexical features described in public IELTS writing descriptors

Band	Accuracy	Range	Sophistication	Error gravity	Appropriacy	Formulaic sequences
9	✓	✓	✓	✓	✓	X
8	✓	✓	✓	✓	✓	✓
7	✓	✓	✓	X	✓	✓
6	✓	✓	✓	✓	✓	X
5	✓	✓	X	✓	✓	X
4	✓	✓	✓	✓	✓	X

From these data, we see that most of the dimensions are included in the descriptors at nearly every band level. These relate to both lexical breadth (range, sophistication) and depth (accuracy, error gravity, appropriacy). However, consideration of formulaic sequences, including collocations, is lacking with 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, there is a danger that raters overlook or undervalue the many types of lexical items. As seen in Chapter 6, collocational sophistication is a significant factor and is therefore deserving of explicit recognition. The wording of rubrics is important, and more experienced raters in fact 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.

As mentioned in Section 3.2.3, a more radical redesign of rating scales (in this case for speaking) has been proposed by Römer (2017) based on the inseparability of grammar and lexis from corpus- and usage-based perspective. To do so, a single 'Lexicogrammatical Resource' category would be formed, as in Ruegg et al. (2011), with explicit inclusion of different types of phraseological elements. However, Römer (2017) herself acknowledges that, historically, language testing researchers (e.g., Lado, 1961) recommended assessing grammar and vocabulary as distinct components, and this practice has continued to the present day. Major international proficiency tests follow this practice, including IELTS and TOEFL iBT, and this practice is mirrored in influential models of language such as Bachman and Palmer (1996, 2010). Therefore, a wholesale redesign of assessment scales is unlikely to be palatable to the teaching and testing communities at the current time. Instead, more minor changes to existing practices could be proposed or implemented which will still have the intended effect of "[acknowledging] the intersection of grammar and vocabulary more explicitly" (Römer, 2017, p. 488), as described in the following section.

7.2.2 Rater training implications

If the goal is to accurately assess impactful elements of lexis in writing, it is necessary to consider more than just the contents of the scales. It is also imperative to reflect on how raters are trained to use them since both training and descriptors have washback effects on the raters (Xerri & Vella Briffa, 2018). As we saw from the qualitative findings in Section 6.2.2, the band descriptors appear to not always match rater beliefs, especially in terms of FSs and accuracy. These considerations therefore fall under the umbrella of *assessment literacy*. Unfortunately, according to most accounts, teachers in many contexts are not provided with sufficient assessment literacy training (Bellhouse, 2018; Boyd & Donnarumma, 2018). However, when teachers are provided with training, spending time on unheeded elements, e.g., formulaic sequences or lexical frequency, can increase construct validity (Fritz & Ruegg, 2013). What is more, teacher education can inform teacher cognition (Borg, 2003), thereby closing the gap between the intentions of assessment bodies and the beliefs of the raters carrying out the assessments.

Assessment bodies that design scales can also benefit from taking into account the beliefs of teachers, i.e., assessment design does not need to be a one-way street. Teachers sometimes perceive high-stakes testing as "divorced from the reality of the classroom" (Xerri & Vella Briffa, 2018, p. 1), but this need not be the case. By consulting and integrating teacher experience and opinions, assessment scales can be made more valid and also better received by the users (Holzknecht et al., 2018). The two descriptors below (Table 16) are 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, the research findings supporting the importance of collocational proficiency, and the current wording of the public descriptors. In doing so, it is

intended to more clearly highlight a key element of lexical proficiency which at present is not given sufficient weight in the analytic scales.

Table 16 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	• make some errors in spelling, word formation, **and collocation**, but they do not impede communication

One challenge of writing descriptors is balancing specificity and practicality as there is very limited space available. In the IELTS descriptors there are usually three bullet points for each band descriptor, and this practice has been maintained in the amended version. Instead of adding new bullet points relating to FSs, minimal alterations have been made (emphasized in bold) to make salient that FS use is part of the existing descriptors for range, sophistication, and accuracy. Note, here the terms *low-*, *mid-*, and *high-frequency* have not been used, maintain the existing more descriptive terms. As suggested previously, in training these terms could be defined more specifically with reference to frequency bands. Another terminological decision was to use the term *multiword expression.* Although this study has used *formulaic sequence* and this term would make sense in these descriptors, *multiword expression* is at this point probably more widely used in the teaching community and more immediately accessible.

7.3 Conclusions

In this dissertation, we began with an overview of research related to lexical proficiency, collocational proficiency, and studies comparing human judgements to statistical measures. Situated in this context, a validation study was presented with the purpose of creating a set of three modified student essays with normalized text lengths which could be used for experimental purposes. This work then culminated in the main study, an investigation of expert IELTS examiner ratings of 30 different versions of the normalized texts which had been manipulated in terms of their collocational sophistication and accuracy. These data were used to answer the three original research questions, summarized here:

RQ1: To what extent do statistical measures of collocational proficiency correlate with and predict expert rater judgments of overall and lexical proficiency in terms of collocational sophistication and collocation accuracy?

Using linear regression models to predict the Lexical Resource fair scores, lexical sophistication was seen to be a significant factor. Furthermore, whether sophisticated words were part of a collocation significantly impacted the ratings. These findings suggest that sophisticated collocations, as operationalized using a collocate-frequency approach, are slightly more impactful on raters' perceptions of text quality as compared to sophisticated words not part of collocations (but still in the context of a text). However, collocational accuracy was not significant in the models, potentially due to the quantity of collocational errors which differentiated the high- and low-accuracy text versions, or perhaps due to the low error gravity of the collocations which did not greatly distort meaning.

RQ2: Is the distinction between, 'low', 'mid', and 'high' frequency lexical items as part of collocations significant?

Because sophistication was significant, it was possible to use post-hoc analyses to see the relevant importance of low-sophistication (K1-2), mid-sophistication (K3-9), and high-sophistication (K10-16) lexical items. These findings indicated that there was a significant difference between low and mid, a greater difference between low and high, but no difference was found between mid and high. Taken together, these results paint a picture of a scalar impact of frequency-based sophistication, rather than a clear categorical (three-way) division.

RQ3: To what extent do expert raters' scores align with the features of the texts that they stated as being most salient/impactful?

The examiners' comments revealed that they noticed and were aware of specific lexical items at both the word and collocation levels, many of which were the experimentally manipulated items. In discussing these items, raters showed a strong focus on sophistication, accuracy, range, and appropriacy, which is unsurprising considering that these elements are consistently present in the band descriptors. However, they also routinely described use of collocations and formulaic sequences which are not explicitly mentioned in the majority of band descriptors. Overall, the examiners' ratings aligned with the statistical sophistication measures, indicating that frequency was an important consideration. However, although accuracy was a primary concern for the examiners, this lexical aspect did not translate to significant rating differences in this experimental framework. This research can therefore be said to have a found a relationship between rater cognition and text features that are *not* mentioned in the rubric.

The main contributions of this study are threefold. From a methodological standpoint, the careful text selection and normalization described in Chapter 5 provides a model for future

research. By carefully normalizing text length and validating the results, the student essays can be used as 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 is uncommon in this field of research but shows merit in terms of accounting for individual rater variability.

The second contribution of this study is to classroom pedagogy for the teaching of lexis. Historically, the teaching of FSs 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). As Vilkaitė-Lozdienė and Schmitt (2020, p. 81) describe it, a central issue of language learning is therefore the following:

> If a person wants to acquire a second language, learning its vocabulary is definitely an important task. However, we cannot teach or learn all the words in a language, as there are simply too many of them. Therefore, some decisions need to be made, and some words have to be prioritized.

The results of this study have suggested that for students to improve the quality of their written academic English,[10] it is beneficial to judiciously include *very* high-sophistication lexis, *even if* learning such lexis is of lesser benefit to developing receptive skills. This call for action should be welcomed by most teachers who feel that learners need to learn more vocabulary than they are learning at present (Bulté et al., 2008).

[10] Here, there is an assumption that "highly rated" is equivalent to "high quality" based on the construct validity of IELTS.

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

7.3.1 Limitations of the study

Many of the limitations of this study are due to conscious decisions regarding its methodological design. The experimental nature of the study required controlling for features such as text length and 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, based on the findings, the potential of other metrics was revealed. Of course, it is impractical to include too many variables as the number of text versions balloons as a result. However, the exclusion of error gravity as a factor somewhat limits the conclusions that can be drawn about the collocational accuracy findings. Finally, as in all such studies, an even greater number of raters would have provided additional useful datapoints about each of the text versions.

7.3.2 Recommendations for future research

Given the limitations listed above, future research might include partial replications of this study but with adjustments to the texts, e.g., to increase the difference in collocation errors or error severity between the high and low accuracy text versions. Adjusting collocation sophistication using a collocation frequency approach would also be informative with respect to types of sophistication. Such an approach might mean varying the morpho-syntax in terms of sophistication of structure vs. the frequency bands of words in those structures. If possible, a similar study but using speech data, e.g., IELTS long-turn speaking tasks, would also reveal whether the trends from this study are also true of spoken output.

The qualitative element of this research, the raters' comments, could also be further explored. For example, raters' cognition and lexical awareness could be investigated, potentially through interviews or surveys to acquire a more thorough understanding of the raters' thought processes and beliefs about lexis and assessment. Teacher beliefs are a primary driver of pedagogical classroom decisions (Gao & Ma, 2011), which is why it is imperative to identify and address mismatches between teacher beliefs and teaching/assessment expectations. Goh and Ang-Aw (2018) provide a useful for template for what such research might entail as they explored teachers' beliefs about oral proficiency using a Think Aloud Protocol and subsequently a delayed beliefs questionnaire. Fifteen years ago, Borg (2006) noted that there had so far been limited research on teacher cognition of vocabulary learning and to date, this exciting area continues to be under-researched (see, e.g., Rossiter et al., 2016). Through projects such as the current study and others in a similar vein, it is 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 an expert audience.

Appendix A Studies comparing statistical features of texts to proficiency ratings

Table 17 Overview of comparison studies

Study	language focus	task type	prompt topic(s)	# of prompts	learner L1	scale types	# of raters	# of learners	# of texts	ratings per text	time limit
Agustin Llach (2007)	lexical errors	letter	self-introduction	1	Spanish	holistic analytic	3	71	71	2-3	30 min
Aryadoust & Liu (2015)	text complexity	expository essay persuasive essay	the internet teens working	2	Chinese	analytic	4	163	326	4	none
Bestgen & Granger (2014)	bigrams	descriptive essay	describing personal life	8	various	analytic	2	57	171	2	30 min
Bulté & Housen (2014)	lexical and syntactical complexity	descriptive essay	describing personal life	8	various	analytic	2	45	90	2	30 min
Crossley & McNamara (2012)	cohesion. sophistication	letter or expository essay	various	4	Chinese	holistic	?	514	514	?	75 min
Crossley et al. (2014)	linguistic microfeatures	TOEFL essays	?	2	various	holistic	3	240	480	1-2	30 min
Dabbagh & Janebi Enayat (2019)	vocabulary breadth/depth	descriptive essay	dream home person they admire	2	Farsi	analytic	2	67	67	2	?
Daller & Phelan (2007)	lexical richness	descriptive essay	home country	1	various	holistic analytic	4	31	31	4	?
Engber (1995)	lexical proficiency	expository essay	studying in US	1	various	holistic	10	66	66	10	35 min
Ferris (1994)	lexical and syntactic features	expository essay	culture shock	1	various	holistic	3	160	160	3	35 min
Fritz & Ruegg (2013)	lexical accuracy, sophistication, range	argumentative essay	vegetarianism	1	Japanese	analytic	27	27	27	2	30 min
Ginther & Grant (1997)	grammatical error	expository essay	teachers	1	various	holistic	2	180	180	2	30 min
Granger & Bestgen (2014)	bigrams	argumentative essay	?	?	various	holistic analytic	2	223	223	223	?

Grant & Ginther (2000)	linguistic features	expository essay	news sources	1	various	holistic	2	90	90	2	30 min
Guo et al. (2013)	essay quality	expository essay	cooperation	1	various	holistic	2	240	240	2	30 min
Jarvis (2013b)	lexical diversity	film description	film description	1	Finnish/ Swedish	holistic	2	210	210	2	?
Jiang et al. (2021)	phraseological complexity	narrative essays	various	4	Chinese	analytic	3	322	322	2-3	30 min
Kyle & Crossley (2015)	lexical sophistication	free writing	NA	NA	various	holistic	?	10	180 244	?	?
Kyle & Crossley (2016)	lexical sophistication	argumentative essay	various	4	various	holistic	3	480	480	2-3	30 min
Kyle & Crossley (2017)	syntactic sophistication	argumentative essay	various	2	various	holistic	3	480	480	2-3	30 min
Kyle et al. (2020)	lexical diversity	argumentative essay	selecting majors cooperation	2	various	holistic	3	300	300	2-3	30 min
Lenko-Szymanska (2019)	lexical proficiency	argumentative essay	mobile phones	1	Polish	holistic analytic	4	150	150	4	90 min
Linnarud (1986)	lexical proficiency	narrative essays	picture prompts	1	Swedish	holistic	15	42	42	15	40 min
Monteiro et al. (2020)	lexical frequency contextual diversity	argumentative essay	selecting majors cooperation	2	various	holistic	3	480	480	2-3	30 min
Ruegg et al. (2011)	lexical qualities	expository essay	methods of communication	1	Japanese	analytic	45	140	140	2	?
Vögelin et al. (2019)	lexical features	argumentative essay	technology	1	German	holistic analytic	37	8	8	?	90 min
Yu (2010)	lexical diversity	?	various	5	various	holistic	?	200	200	2	30 min

Appendix B Original texts and examiner comments

Appendix B.1 Level B1 original text and comments

I disagree that point about children brought up in families.
because, I show that situation arounds me at our country parents.
They want they had everything give to their children.
but, their behavior is not good effect to them
On the other hands, children brought up by wealthy parents.
they are strong, that means they can do prepare to deal with
the problems of adult life
In my case, I start work from 20 ages I had social experience
and I got a money for myself. however, My age is late to work
by children ages and I heard about child doing work by another
countries that countries had a culture about childrens
That is, they doing work for their pocketmoney.
they could their money buy something or entrance to the bank.
also, our country childrens do this. but, many childrens accept the
money by their parents. which persons got a pocketmoney over the 20 ages
I think, if childrens had a work and they study at money,
they perfectly prepared their adult life after they must be parents

Examiner comment

Band 4

While it is obviously related to the topic, the introduction is confusing and the test taker's position is difficult to identify. Ideas are limited and although the test taker attempts to support them with examples from experience, they remain unclear. There is no overall progression in the response and the ideas are not coherently linked. Although cohesive devices are used, they assist only minimally in achieving coherence. The range of vocabulary is basic and control is inadequate for the task. Language from the input material is used inappropriately and frequent errors in word choice and collocation cause severe problems for the reader. Similarly, the range of structures is very limited, the density of grammatical and punctuation error is high and these features cause some difficulty for the reader. Attempts to use complex structures, such as subordination, are rare and tend to be very inaccurate.

Figure 9 Level B1 original text and comments (IELTS, n.d.-d)

Appendix B.2 Level B2 original text and comments

I greatly support the idea about Children who are brought up in families that do not have large amounts of money are better prepared to deal with the problems of adult life than children brought up by wealthy parents. I support it, because of the following reason.

Children who are brought up in families that do not have large amounts of money are raise in a certain psychological values. Such as the value of hardwork, dicipline, they are used to be in the condition where money doesn't come easily. They have to earn it, work for it. Oppose to it, a child who comes from a wealthy family is used to have money all the time. whenever they want something, the money is easilly gave to them as if everyday are their birthday.

Children who are brought up in families that do not have large amounts of money are well-trained to face adulthood. They are well-prepared to see the fact that the world is a very tough place. They watched their parent everyday worked very hard just to put food on the table. They have the advantage to see the reality and embrace it, set their mind that they too have work hard for their future, their own dreams, their authentic self. A child that came from a wealthy family doesn't always have this advantage. This is because their eyes are blinded by the power of money, that their parent has. They also have a disadvantage of a family love life. Commonlly

wealthy parents express love by money. They love their children, so they bought them cars, expensive clothes, toys, but they are never home when their children needs them. The basic necessity of compassion isn't fulfilled in this kind of family. The impact to a child is that they will grow up and think that money is everything, that the source of happiness is money. They don't care about other people, they only care about money. The problem is they don't know how to get it, they've been spoiled all the time, so doesn't have the time to ~~discover~~ discovered the art of money making, only money spending. On the contrary children from families that do not have large amount of money will grow up with the sense of respect for money, they know how to get it and use it well. They know how to face adult life problems because they've been watching since they were a child. But a wealthy child is always to busy with himself to know that.

Examiner comment

Band 6.5

The introduction is mainly copied from the rubric. The arguments are generally well developed and there is a clear position, despite the lack of a conclusion. Better use of paragraphing would have allowed a clearer focus to some of the supporting points and prevented the lapse into generalisation towards the end. Nevertheless, there is a generally clear progression with a good arrangement of opposing arguments. Referencing is usually accurate and effective, but better use of linkers would have improved the cohesion. Vocabulary is sufficient and used with some flexibility. The choice is not always precise but the test taker can evidently incorporate less common/idiomatic phrases into the argument and there is a good range that is generally accurate. The repetition of language from the rubric, while integrated, reveals a lack of ability to paraphrase. Regular errors detract from the use of a range of structures, although they do not impede communication. This is a generally good response to the task, but the weaknesses in organisation and grammatical control limit the rating to Band 6.5.

Figure 10 Level B2 original text and comments (IELTS, n.d.-d)

Appendix B.3 Level C1 original text and comments

I do agree to the statement that children brought up in poor families are better prepared to deal with the problems of adult life than children brought up ~~by~~ by wealthy parents.

Children of poor parents are preuaturely exposed to the problems of adult life eg. ~~earning a living~~ learning to survive on a low family income and sacrificing luxuries for essential items. These children begin to see the 'realities' of life in their home or social environment. Their parents own struggles serve as an example to them.

These children are taught necessary skills for survival as an adult from a very early age. Many children eg work in the weekends or holidays to either collect some pocket money or even contribute to their families' income. A good example is the many children who accompany their parents to sell produce at the market. They are making a direct contribution to their families in terms of labor or income.

Children of poor families also are highly motivated. They tend to set high goals to improve their economic & social situation. A relevant example would be Mr Bill Gates (founder of Microsoft Corporation) He had an impoverished background but he used his talent and motivation to set up the worlds largest computer organisation.

However, there are some problems that children from poor backgrounds do encounter. Many of these children who are 'robbed' of their childhood eg while working, may feel cheated. They often turn to crime This however, is a small group.

In summing up, children with impoverished backgrounds are able to deal with problems of adult life because of early exposure, family role models and sheer motivation.

Examiner comment

Band 8

The topic is very well addressed and the position is clear throughout. Main ideas are presented and well supported, apart from some over-generalisation in the penultimate paragraph. The ideas and information are very well organised and paragraphing is used appropriately throughout. The answer can be read with ease due to the sophisticated handling of cohesive devices – only the lack of an appropriate introduction and the minor error in the second use of 'eg' mars this aspect of the response. The writer uses a wide and very natural range of vocabulary with full flexibility. There are many examples of appropriate modification, collocation and precise vocabulary choice. Syntax is equally varied and sophisticated. There are only occasional errors in an otherwise very accurate answer. Overall this performance is a good example of Band 8.

Figure 11 Level C1 original text and comments (IELTS, n.d.-d)

Appendix C IELTS scores and CEFR levels

Table 18 IELTS band scores and descriptors (IELTS, n.d.-b)

Band score	Skill level	Description
9	Expert user	The test taker has fully operational command of the language. Their use of English is appropriate, accurate and fluent, and shows complete understanding.
8	Very good user	The test taker has fully operational command of the language with only occasional unsystematic inaccuracies and inappropriate usage. They may misunderstand some things in unfamiliar situations. They handle complex and detailed argumentation well.
7	Good user	The test taker has operational command of the language, though with occasional inaccuracies, inappropriate usage and misunderstandings in some situations. They generally handle complex language well and understand detailed reasoning.
6	Competent user	The test taker has an effective command of the language despite some inaccuracies, inappropriate usage and misunderstandings. They can use and understand fairly complex language, particularly in familiar situations.
5	Modest user	The test taker has a partial command of the language and copes with overall meaning in most situations, although they are likely to make many mistakes. They should be able to handle basic communication in their own field.
4	Limited user	The test taker's basic competence is limited to familiar situations. They frequently show problems in understanding and expression. They are not able to use complex language.
3	Extremely limited user	The test taker conveys and understands only general meaning in very familiar situations. There are frequent breakdowns in communication.
2	Intermittent user	The test taker has great difficulty understanding spoken and written English.
1	Non-user	The test taker has no ability to use the language except a few isolated words.
0	Did not attempt the test	The test taker did not answer the questions.

Table 19 CEFR levels and descriptors (Council of Europe, 2021)

PROFICIENT USER	C2	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
	C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.
INDEPENDENT USER	B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes & ambitions and briefly give reasons and explanations for opinions and plans.
BASIC USER	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

Appendix D Length-normalized texts

Legend

Text copied from rubric

Added text

Changed text

~~Removed text~~

Accurate collocations

Inaccurate collocations

Appendix D.1 Level B1 base text

I disagree that point about children brought up in families are prepared their life and after are good parents. because, I show that situation around me at our country parents.
They want they had everything give to their children and could buy many things like good school. but, their behavior is not good effect to them
On the other hands, children brought up by wealthy parents, they are strong, that means they can do prepare to deal with the problems of adult life. They work for having money that could buy for everything they want
In my case, I start work from 20 ages. I start work my country and work as a journalist for money. I had social experience and I got a money for myself. however, My age is late to work by children ages and I heard about child doing work by another countries that countries had a culture about children. They start work when they 15, and it is very young. ~~That is~~ They doing work for their pocket money that is good. they could their money buy something or entrance to the bank. That is good they could do buy something. for their school and their parents.
also, our country children's do this but, many children accept the money by their parents which persons got a pocket money over the 20 ages. But I very disagree that point. I think, if children had a work and they study at money, they perfectly prepared their adult life after they must be parents.

Appendix D.2 Level B2 base text

I greatly support the idea about children who are brought up in families that do not have large amounts of money are better prepared to deal with the problems of adult life than children brought up by wealthy parents~~. I support it, because of the following reason.~~

Children who are brought up in families that do not have large amounts of money are raise in a certain psychological values. Such as the value of hard work, discipline, they are used to be in the condition where money doesn't come easily. ~~They have to earn it, work for it.~~ Oppose to it, a child who comes from a wealthy family is used to have money all the time. Whenever they want something, the money is fast given to them ~~as if every day are their birthday~~.

Children who are brought up in families that do not have large amounts of money are skilled to face adulthood. ~~They are well-prepared to see the fact that the world is a very tough place.~~ They watched their parent everyday worked very hard just to put food in the table. They have the advantage to see the reality and embrace it, set their mind that they too have work hard for their future~~, their own dreams, their authentic self~~. A child that came from a wealthy family doesn't always have the advantage~~. This is~~ because their eyes are blinded by the power of money, that their parent has. They also have a disadvantage of a family love life. Commonly wealthy parents express love by money~~. They love their children,~~ so they got them cars, nice clothes, toys, but they are never home ~~when their children need them. The basic necessity of compassion isn't fulfilled in this kind of family.~~ The impact to a child is that they will grow up and think that money is everything, that the source of happiness is money. ~~They don't care about other people, they only care about money.~~ The problem is they don't know how to get it, they've been spoiled all the time, so doesn't have the time to discover~~ed~~ the art of money making, only money spending. On the contrary, children from families that do not have large amount of money will grow up with the sense of respect for money, they know how to ~~get it and use it well. They know how to~~ face adult life problems because they've been observing since they were a child. ~~But a wealthy child is always too busy with himself to know that.~~

Appendix D.3 Level C1 base text

I do agree to the statement that children brought up in poor families are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

Children of poor parents are prematurely exposed to the problems of adult life e.g. learning to survive on a low family income and sacrificing luxuries for essential items. These children began to see the 'realities' of life in their home or social environment. Their parents own struggles serve as an example to them.

These children are taught necessary skills for survival as an adult from a very early age. Many children eg work in the weekends or holidays to either collect some pocket money or even contribute to their families' income. A good example is the many children who accompany their parents to sell produce at the market. They are making a straight contribution to their families in terms of labor or income.

Children of poor families also are highly motivated. They tend to set high goals to improve their economic & social situation. A relevant example would be Mr. Bill Gates (founder of Microsoft Corporation) He had an impoverished background but he used his talent and motivation to set up the world's largest computer organization.

However, there are some problems that children from poor backgrounds do encounter. Many of these children who are 'robbed' of their childhood eg while working, may feel cheated. They often turn to crime. This however, is a small group.

In summing up, children with impoverished backgrounds are able to deal with problems of adult life because of early exposure, family role models and determined motivation

Appendix E Collocation identification checklist

Essential criteria

By my judgement…

1. the wordstring consists of 2-3 lexical words (nouns, verbs, adjectives, adverbs, or prepositions) and potentially other grammatical words (e.g., pronouns or determiners).
2. the wordstring is not a phrasal verb (e.g., *brought up*), a compound noun (e.g., *pocket money*), or a proper noun (e.g., *Microsoft corporation*).
3. the wordstring is a community-wide formula for ESL/EFL language teachers and learners, i.e., it constitutes a chunk.
4. there is a greater than chance-level probability that the writer will have encountered this precise wordstring before, from other spoken or written texts.

Guiding criteria

By my judgement…

5. this wordstring has a cohesive meaning or function as a phrase and may lack semantic transparency.
6. this wordstring is associated with a specific situation and/or register.
7. this precise formulation is the one most commonly used by this writer when conveying this idea.
8. this wordstring has been marked grammatically, lexically, or with punctuation in a way that gives it special status as a unit.
9. this wordstring is formulaic, but it has been unintentionally applied inappropriately.
10. this wordstring has greater complexity than other output in the text.
11. this wordstring is more likely to be worth teaching as a bona fide collocation.

* Items 1, 2, 6, 9 are based in part on criteria from Coulmas (1979)
* Items 4-9 are based in part on criteria from Wray and Namba (2003) cited in
* Item 10 is based in part on criteria from Wood (2002)
* Items 2, 5, 11 are based in part on Simpson-Vlach and Ellis (2010)

Appendix F Text versions

Appendix F.1 Level B1 text versions

Text ID: 1

I disagree that point about children brought up in families are prepared their life and then are good parents. because, I show that situation around me at our country parents. They want they had everything give to their children and also could buying things like positive school. but, their behavior is not good effect to them
On the other hand, children brought up by wealthy parents, they are strong, that means they can be prepare with many problems of being adults. They working to have money, that could buy everything they want
In my case, I start work from 20 ages I start work My country and work as a journalist for money I had social experience and I got a money for myself. however, My age is late to work by children ages and I heard about children doing work by another countries that countries had a culture about children They start work when they 15, and it is very young. They doing work for their pocket money and is good
they could their money buy something or entrance to the bank. that is good they could buy something. now and for future time. also, our country's children do this. but, many children accept the money by their parents. which persons got a pocket money over the 20 ages. but I am very disagree that point. I think, if children had to work and they study at money, they perfectly prepared their adult life after they must be parents.

Text ID: 2

I disagree that point about children brought up in families are prepared their life and then are good parents. because, I see that situation around me at our country parents. They want they had everything give to their children and also could buying things like excellent school. but, their behavior is not good effect on them
On the other hand, children brought up by wealthy parents, they are strong, that means they can be prepare with many problems of being adults. They working to have money, that could buy everything they want
In my case, I start work from 20 ages I start work My country and work as a journalist for money I had life experience and I got a money for myself. however, My age is late to work by children ages and I heard about children doing work by another countries that countries had a culture about children They start work when they 15, and it is very young. They doing work for their pocket money and is good
they could their money buy something or entrance to the bank. that is good they could buy something. now and for the future. also, our country's children do this. but, many children accept the money by their parents. which persons got a pocket money over the 20 ages. but I must really disagree with that. I think, if children had to work and they study at money, they perfectly prepared their adult life after they must be parents.

Text ID: 3

I disagree that point about children brought up in families are prepared their life and then are tremendous parents. because, I show that situation around me at our country parents. They want they had everything give to their children and also could buying things like positive school. but, their behavior is not desired effect to them
On the flip side, children brought up by wealthy parents, they are strong, that means they can be prepare with many problems of being adults. They working to acquire money, that could buy everything they want
For me personally, I work independently from 20 ages I start work My country and work as a journalist for money I had social experience and I got a money for myself. however, My age is late to work by children ages and I surprisingly saw children doing chores by another countries that countries had a culture about children They start work when they 15, and it is fairly young. They doing work for their pocket money and is good
they could their money buy something or entrance to the bank. that is good they could buy something. now and for future time. also, our country's children do this. but, many children accept the money by their parents. which persons got a pocket money over the 20 ages. but I am very disagree with that. I think, if children had to work and they study at money, they perfectly prepared their adult life after they must be parents.

Text ID: 4

I disagree that point about children brought up in families are prepared their life and then are tremendous parents. because, I see that situation around me at our country parents. They want they had everything give to their children and also could buying things like excellent school. but, their behavior is not desired effect on them
On the flip side, children brought up by wealthy parents, they are strong, that means they can be prepare with many problems of being adults. They working to acquire money, that could buy everything they want
For me personally, I work independently from 20 ages I start work My country and work as a journalist for money I had life experience and I got a money for myself. however, My age is late to work by children ages and I heard about children doing chores by another countries that countries had a culture about children They start work when they 15, and it is fairly young. They doing work for their pocket money and is good
they could their money buy something or entrance to the bank. that is good they could buy something. now and for the future. also, our country's children do this. but, many children accept the money by their parents. which persons got a pocket money over the 20 ages. but I must really disagree with that. I think, if children had to work and they study at money, they perfectly prepared their adult life after they must be parents.

Text ID: 5

I disagree that point about children brought up in families are prepared their life and then are good parents. because, I show that situation around me at our country parents. They want they had everything give to their children and also could purchasing things like positive school. but, their behavior is not good effect to them
On the other hand, children brought up by wealthy parents, they are mature, that means they can be prepare with many problems of being adults. They working to have money, that could buy everything they want
In my case, I start work from 20 ages I start work My country and work as a journalist for wages I had social experience and I got a money for myself. nevertheless, My age is late to work by children ages and I heard about children doing work by another countries that countries had a culture about children They start work when they 15, and it is very young. They doing work for their pocket money and is good
they could their money buy something or entrance to the bank. that is good they could buy something. now and for future time. also, our tribe's children do this. but, many children accept the money by their parents. which teens got a pocket money over the 20 ages. but I am very disagree with that. I suspect, if children had to work and they study at money, they perfectly prepared their adult life subsequently they must be content.

Text ID: 6

I disagree that point about children brought up in families are prepared their life and then are good parents. because, I see that situation around me at our country parents. They want they had everything give to their children and also could purchasing things like excellent school. but, their behavior is not good effect on them
On the other hand, children brought up by wealthy parents, they are mature, that means they can be prepare with many problems of being adults. They working to have money, that could buy everything they want
In my case, I start work from 20 ages I start work My country and work as a journalist for wages I had life experience and I got a money for myself. nevertheless, My age is late to work by children ages and I heard about children doing work by another countries that countries had a culture about children They start work when they 15, and it is very young. They doing work for their pocket money and is good
they could their money buy something or entrance to the bank. that is good they could buy something. now and for the future. also, our tribe's children do this. but, many children accept the money by their parents. which teens got a pocket money over the 20 ages. but I must really disagree with that. I suspect, if children had to work and they study at money, they perfectly prepared their adult life subsequently they must be content.

Text ID: 7

I disagree that point about children brought up in families are prepared their life and then are first-class parents. because, I show that situation around me at our country parents. They want they had everything give to their children and also could buying things like positive school. but, their behavior is not anticipated effect to them
To corroborate this claim, children brought up by wealthy parents, they are strong, that means they can be prepare with many problems of being adults. They working to hoard money, that could buy everything they want
Not to generalize, I start freelance from 20 ages I start work My country and work as a receptionist for money I had social experience and I got a money for myself. however, My age is late to work by children ages and I saw firsthand children doing work by another countries that countries had a culture about children They start work when they 15, and it is comparatively young. They doing errands for their pocket money and is good
they could their money buy something or safeguard for the future. that is good they could buy something. now and for future time. also, our country's children do this. but, many children accept the money by their parents. which persons got a pocket money over the 20 ages. but I am very disagree with that. I think, if children had to work and they study at money, they psychologically prepared their adult life after they must be parents.

Text ID: 8

I disagree that point about children brought up in families are prepared their life and then are first-class parents. because, I see that situation around me at our country parents. They want they had everything give to their children and also could buying things like excellent school. but, their behavior is not anticipated effect on them
To corroborate this claim, children brought up by wealthy parents, they are strong, that means they can be prepare with many problems of being adults. They working to hoard money, that could buy everything they want
Not to generalize, I start freelance from 20 ages I start work My country and work as a receptionist for money I had life experience and I got a money for myself. however, My age is late to work by children ages and I saw firsthand children doing work by another countries that countries had a culture about children They start work when they 15, and it is comparatively young. They doing errands for their pocket money and is good
they could their money buy something or safeguard for the future. that is good they could buy something. now and for the future. also, our country's children do this. but, many children accept the money by their parents. which persons got a pocket money over the 20 ages. but I must really disagree with that. I think, if children had to work and they study at money, they psychologically prepared their adult life after they must be parents.

Text ID: 9

I disagree that point about children brought up in families are prepared their life and then are good parents. because, I show that situation around me at our country parents. They fantasize they had everything give to their children and also could procuring things like positive school. but, their behavior is not good effect to them
On the other hand, children brought up by wealthy parents, they are resilient, that means they can be prepare with many problems of being adults. They working to have money, that could buy everything they want
In my case, I start work from 20 ages I start work My country and work as a journalist for tabloids I had social experience and I got a money for myself. unbelievably, My age is late to work by children ages and I heard about children doing work by another countries that countries had a culture about children They start work when they high-school, and it is very young. They doing work for all their pastimes and is good
they could their money buy something or entrance to the bank. that is good they could buy something. now and for future time. also, our locality's children do this. but, many children accept the money by their parents. which juveniles got a pocket money over the 20 ages. but I am very disagree with that. I posit, if children had to work and they study at money, they perfectly prepared their adult life logically they must be grown-up.

Text ID: 10

I disagree that point about children brought up in families are prepared their life and then are good parents. because, I see that situation around me at our country parents. They fantasize they had everything give to their children and also could procuring things like excellent school. but, their behavior is not good effect on them
On the other hand, children brought up by wealthy parents, they are resilient, that means they can be prepare with many problems of being adults. They working to have money, that could buy everything they want
In my case, I start work from 20 ages I start work My country and work as a journalist for tabloids I had life experience and I got a money for myself. unbelievably, My age is late to work by children ages and I heard about children doing work by another countries that countries had a culture about children They start work when they high-school, and it is very young. They doing work for all their pastimes and is good
they could their money buy something or entrance to the bank. that is good they could buy something. now and for the future. also, our locality's children do this. but, many children accept the money by their parents. which juveniles got a pocket money over the 20 ages. but I must really disagree with that. I posit, if children had to work and they study at money, they perfectly prepared their adult life logically they must be grown-up.

Appendix F.2 Level B2 text versions

Text ID: 11

I greatly support the idea about children who are brought up in families that do not have large amounts of money are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

Children who are brought up in families that do not have large amounts of money are raised in a certain psychological values. Such as the value of hard work, discipline, they are used to be in the condition where money doesn't come easily. Oppose to it, a child who comes from a wealthy family is used to have money all the time. Whenever they want something, the money is easily gave to them.

Children who are brought up in families that do not have large amounts of money are well-trained to face adulthood. They watched their parent every day worked very hard just to put food in the table. They have the advantage to see the reality and embrace it, set their mind that they too have work hard for their future. A child that came from a wealthy family doesn't always have the advantage. This is because their eyes are blinded by the power of money, that their parent has. They also have a disadvantage of a family love life. Commonly wealthy parents express love by money, so they bought them cars, expensive clothes, toys, but they are never home. The impact to a child is that they will grow up and think that money is everything, that the source of happiness is money. The obstacle is they don't know how to get it, they've been spoiled in every time, so doesn't have the time to discovered the art of money making, only money spending. On the contrary, children from families that do not have large amount of money will grow up with the sense of respect for money, they do know how to face adult life problems because they've been observing since they were a child.

Text ID: 12

I greatly support the idea about children who are brought up in families that do not have large amounts of money are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

Children who are brought up in families that do not have large amounts of money are raised in a certain ethical values. Such as the value of hard work, discipline, they are used to be in the condition where money doesn't come easily. Oppose to it, a child who comes from a wealthy family is used to have money all the time. Whenever they want something, the money is easily gave to them.

Children who are brought up in families that do not have large amounts of money are equipped to face adulthood. They watched their parent every day worked very hard just to put food on the table. They have the advantage to see the reality and embrace it, decide for themselves that they too have work hard for their future. A child that came from a wealthy family doesn't always have the advantage. This is because their eyes are blinded by the power of money, that their parent has. They also have a disadvantage of a happy home life. Commonly wealthy parents express love through money, so they bought them cars, expensive clothes, toys, but they are never home. The impact to a child is that they will grow up and think that money is everything, that the source of happiness is money. The obstacle is they don't know how to get it, they've been spoiled in every time, so doesn't have the time to discovered the art of money making, only money spending. On the contrary, children from families that do not have large amount of money will grow up with the sense of respect for money, they do know how to face adult life problems because they've been observing since they were a child.

Text ID: 13

I greatly endorse the idea about children who are brought up in families that do not have large amounts of money are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

Children who are brought up in families that do not have large amounts of money are raised in a certain psychological values. Such as the virtue of hard work, discipline, they are used to be in the condition where money isn't readily obtained. Oppose to it, a child who comes from a wealthy family is used to possess wealth without a doubt. Whenever they want something, the money is easily gave to them.

Children who are brought up in families that do not have large amounts of money are well-trained to face adulthood. They watched their parent every day worked incredibly hard just to put food in the table. They have the advantage to see the reality and embrace it, set their mind that they too have work hard for their future. A child that came from a wealthy family doesn't always have the advantage. This is because their eyes are blinded by the power of money, that their parent has. They also have a disadvantage of a family love life. Commonly wealthy parents express love by money, so they bought them cars, designer clothes, toys, but they are seldom home. The impact to a child is that they will grow up and think that money is everything, that the source of happiness is money. The obstacle is they don't know how to get it, they've been spoiled in every time, so doesn't have the time to discovered the necessity for money making, only money spending. On the contrary, children from families that do not have large amount of money will be impressed with the sense of gratitude for money, they do know how to face adult life problems because they've been observing since they were a child.

Text ID: 14

I greatly endorse the idea about children who are brought up in families that do not have large amounts of money are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

Children who are brought up in families that do not have large amounts of money are raised in a certain ethical values. Such as the virtue of hard work, discipline, they are used to be in the condition where money isn't readily obtained. Oppose to it, a child who comes from a wealthy family is used to possess wealth without a doubt. Whenever they want something, the money is easily gave to them.

Children who are brought up in families that do not have large amounts of money are equipped to face adulthood. They watched their parent every day worked incredibly hard just to put food on the table. They have the advantage to see the reality and embrace it, decide for themselves that they too have work hard for their future. A child that came from a wealthy family doesn't always have the advantage. This is because their eyes are blinded by the power of money, that their parent has. They also have a disadvantage of a happy home life. Commonly wealthy parents express love through money, so they bought them cars, designer clothes, toys, but they are seldom home. The impact to a child is that they will grow up and think that money is everything, that the source of happiness is money. The obstacle is they don't know how to get it, they've been spoiled in every time, so doesn't have the time to discovered the necessity for money making, only money spending. On the contrary, children from will families that do not have large amount of money be impressed with the sense of gratitude for money, they do know how to face adult life problems because they've been observing since they were a child.

Text ID: 15

I greatly support the idea about children who are brought up in families that do not have large amounts of money are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

Children who are brought up in families that do not have large amounts of money are raised in a certain psychological values. Such as the value of hard work, discipline, they are used to endure in the condition where money doesn't come easily. Oppose to it, a teen who comes from a wealthy family is used to have money all the time. Whenever they want luxuries, the money is easily gave to them.

Children who are brought up in families that do not have large amounts of money are well-trained to face adulthood. They perceived their parent every day worked very hard just to put food in the table. They have the advantage to see the reality and embrace it, set their mind that they likewise have work hard for their future. A child that came from a wealthy family doesn't always have the advantage. This is because their eyes are blinded by the power of money, that their parent possesses. They also have a disadvantage of a family love life. Commonly wealthy parents express love by money, so they bought them electronics, expensive clothes, toys, but they are never home. The impact to a child is that they will grow up and presume that money is vital, that the source of happiness is money. The obstacle is they don't know how to attain it, they've been spoiled in every time, so doesn't have the time to acquired the art of money making, only money spending. On the contrary, children from families that do not have large amount of money will grow up with the sense of respect for prosperity, they do know how to face adult life problems because they've been observing since they were a child.

Text ID: 16

I greatly support the idea about children who are brought up in families that do not have large amounts of money are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

Children who are brought up in families that do not have large amounts of money are raised in a certain ethical values. Such as the value of hard work, discipline, they are used to endure in the condition where money doesn't come easily. Oppose to it, a teen who comes from a wealthy family is used to have money all the time. Whenever they want luxuries, the money is easily gave to them.

Children who are brought up in families that do not have large amounts of money are equipped to face adulthood. They perceived their parent every day worked very hard just to put food on the table. They have the advantage to see the reality and embrace it, decide for themselves that they likewise have work hard for their future. A child that came from a wealthy family doesn't always have the advantage. This is because their eyes are blinded by the power of money, that their parent possesses. They also have a disadvantage of a happy home life. Commonly wealthy parents express love through money, so they bought them electronics, expensive clothes, toys, but they are never home. The impact to a child is that they will grow up and presume that money is vital, that the source of happiness is money. The obstacle is they don't know how to attain it, they've been spoiled in every time, so doesn't have the time to acquired the art of money making, only money spending. On the contrary, children from families that do not have large amount of money will grow up with the sense of respect for prosperity, they do know how to face adult life problems because they've been observing since they were a child.

Text ID: 17

I greatly commend the idea about children who are brought up in families that do not have large amounts of money are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

Children who are brought up in families that do not have large amounts of money are raised in a certain psychological values. Such as the virtues of being diligent, discipline, they are used to be in the condition where money isn't freely bestowed. Oppose to it, a child who comes from a wealthy family is used to unimaginable wealth as a constant. Whenever they want something, the money is easily gave to them.

Children who are brought up in families that do not have large amounts of money are well-trained to face adulthood. They watched their parent every day worked very hard just to put food in the table. They have the advantage to see the reality and embrace it, set their mind that they too have work hard for their future. A child that came from a wealthy family doesn't always have the advantage. This is because their eyes are dazzled by the allure of wealth, that their parent has. They also have a disadvantage of a family love life. Commonly wealthy parents express love by money, so they bought them cars, trendy clothes, toys, but they are never home. The impact to a child is that they will grow up and think that money is everything, that the path to fulfillment is money. The obstacle is they don't know how to get it, they've been spoiled in every time, so doesn't have the time to discovered a knack for money making, only money spending. In startling contrast, children from families that do not have large amount of money will grow up with the sense of respect for money, they do know how to face adult life problems because they've been observing since they were a child.

Text ID: 18

I greatly commend the idea about children who are brought up in families that do not have large amounts of money are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

Children who are brought up in families that do not have large amounts of money are raised in a certain ethical values. Such as the virtues of being diligent, discipline, they are used to be in the condition where money isn't freely bestowed. Oppose to it, a child who comes from a wealthy family is used to unimaginable wealth as a constant. Whenever they want something, the money is easily gave to them.

Children who are brought up in families that do not have large amounts of money are equipped to face adulthood. They watched their parent every day worked very hard just to put food on the table. They have the advantage to see the reality and embrace it, decide for themselves that they too have work hard for their future. A child that came from a wealthy family doesn't always have the advantage. This is because their eyes are dazzled by the allure of wealth, that their parent has. They also have a disadvantage of a happy home life. Commonly wealthy parents express love through money, so they bought them cars, trendy clothes, toys, but they are never home. The impact to a child is that they will grow up and think that money is everything, that the path to fulfillment is money. The obstacle is they don't know how to get it, they've been spoiled in every time, so doesn't have the time to discovered a knack for money making, only money spending. In startling contrast, children from families that do not have large amount of money will grow up with the sense of respect for money, they do know how to face adult life problems because they've been observing since they were a child.

Text ID: 19

I greatly support the idea about children who are brought up in families that do not have large amounts of money are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

Children who are brought up in families that do not have large amounts of money are raised in a certain psychological values. Such as the value of hard work, discipline, they are used to hustle in the condition where money doesn't come easily. Oppose to it, a juvenile who comes from a wealthy family is used to have money all the time. Whenever they want gratification, the money is easily gave to them.

Children who are brought up in families that do not have large amounts of money are well-trained to face adulthood. They scrutinized their parent every day worked very hard just to put food in the table. They have the advantage to see the reality and embrace it, set their mind that they too have work hard for their future. A child that came from a wealthy family doesn't always have the advantage. This is because their eyes are blinded by the power of money, that their parent flaunts. They also have a disadvantage of a family love life. Ordinarily wealthy parents express love by money, so they bought them cars, expensive clothes, belongings, but they are never home. The impact to a child is that they will grow up and surmise that money is paramount, that the source of happiness is money. The obstacle is they don't know how to get it, they've been pampered in every time, so doesn't have the time to instilled the art of money making, only money spending. On the contrary, children from families that do not have large amount of money will grow up with the sense of respect for perseverance, they do know how to face adult life problems because they've been observing since they were a child.

Text ID: 20

I greatly support the idea about children who are brought up in families that do not have large amounts of money are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

Children who are brought up in families that do not have large amounts of money are raised in a certain ethical values. Such as the value of hard work, discipline, they are used to hustle in the condition where money doesn't come easily. Oppose to it, a juvenile who comes from a wealthy family is used to have money all the time. Whenever they want gratification, the money is easily gave to them.

Children who are brought up in families that do not have large amounts of money are equipped to face adulthood. They scrutinized their parent every day worked very hard just to put food on the table. They have the advantage to see the reality and embrace it, decide for themselves that they too have work hard for their future. A child that came from a wealthy family doesn't always have the advantage. This is because their eyes are blinded by the power of money, that their parent flaunts. They also have a disadvantage of a happy home life. Ordinarily wealthy parents express love through money, so they bought them cars, expensive clothes, belongings, but they are never home. The impact to a child is that they will grow up and surmise that money is paramount, that the source of happiness is money. The obstacle is they don't know how to get it, they've been pampered in every time, so doesn't have the time to instilled the art of money making, only money spending. On the contrary, children from families that do not have large amount of money will grow up with the sense of respect for perseverance, they do know how to face adult life problems because they've been observing since they were a child.

Appendix F.3 Level C1 text versions

Text ID: 21

I do agree to the statement that children brought up in poor families are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

Children of poor parents are prematurely exposed to problems of adult life e.g. learning to survive on a low family income and sacrificing luxuries for essential items. These children began to see the 'realities' of life in their home or social environment. Their parent's own struggles serve as an example to them.

These children are taught necessary skills for survival as an adult from a very early age. Many children eg work in the weekends to either collect some pocket money or even contribute to their families' income. A good example is the many children who accompany their parents to sell produce at the market. They are making a straight contribution to their families in terms of labor or income.

Children of poor families also are highly motivated. They tend to set high goals to improve their economic & social situation. A relevant example would be Mr Bill Gates (founder of Microsoft Corporation) He had an impoverished background but he used his talent and motivation to set up the worlds largest computer organization.

However, there are some problems that children from poor backgrounds encounter. Many of these children who are 'robbed' of their childhood eg while working, may feel cheated. They often turn to crime. This however, is a small group.

In summing up, children with impoverished backgrounds are able to deal with problems of adult life because of early exposure, family role models and sheer motivation.

Text ID: 22

I do agree with the statement that children brought up in poor families are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

Children of poor parents are prematurely exposed to problems of adult life e.g. learning to survive on a low family income and sacrificing luxuries for essential items. These children began to see the 'realities' of life in their home or social environment. Their parent's own struggles serve as an example to them.

These children are taught necessary skills for survival as an adult from a very early age. Many children eg work on the weekend to either earn some pocket money or even contribute to their families' income. A good example is the many children who accompany their parents to sell produce at the market. They are making a direct contribution to their families in terms of labor or income.

Children of poor families also are highly motivated. They tend to set high goals to improve their economic & social situation. A relevant example would be Mr Bill Gates (founder of Microsoft Corporation) He had an impoverished background but he used his talent and motivation to set up the worlds largest technology company.

However, there are some problems that children from poor backgrounds encounter. Many of these children who are 'robbed' of their childhood eg while working, may feel cheated. They often turn to crime. This however, is a small group.

To sum up, children with impoverished backgrounds are able to deal with problems of adult life because of early exposure, family role models and sheer motivation.

Text ID: 23

I do agree to the statement that children brought up in poor families are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

Children of humble parents are prematurely exposed to problems of adult life e.g. learning to cope on a low family income and sacrificing luxuries for essential items. These children began to see the 'realities' of life in their home or surrounding environment. Their parent's own struggles act as an incentive to them.

These children are taught fundamental skills for survival as an adult from a very early age. Many children eg work in the weekends to either collect some pocket money or even occasionally supplement their families' income. A concrete example is the many children who accompany their parents to sell produce at the market. They are making a straight contribution to their families with regard to labor or income.

Children of poor families also are truly inspired. They tend to set ambitious goals to upgrade their economic & social status. A relevant example would be Mr Bill Gates (founder of Microsoft Corporation) He had a modest background but he used his talent and motivation to set up the worlds largest computer organization.

However, there are some problems that children from poor backgrounds encounter. Many of these children who are 'robbed' of their childhood eg while working, may feel cheated. They often turn to crime. This however, is a small group.

In summing up, children with modest backgrounds are able to deal with problems of adult life because of early exposure, family role models and sheer motivation.

Text ID: 24

I do agree with the statement that children brought up in poor families are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

Children of humble parents are prematurely exposed to problems of adult life e.g. learning to cope on a low family income and sacrificing luxuries for essential items. These children began to see the 'realities' of life in their home or surrounding environment. Their parent's own struggles act as an incentive to them.

These children are taught fundamental skills for survival as an adult from a very early age. Many children eg work on the weekend to either earn some pocket money or even occasionally supplement their families' income. A concrete example is the many children who accompany their parents to sell produce at the market. They are making a direct contribution to their families with regard to labor or income.

Children of poor families also are truly inspired. They tend to set ambitious goals to upgrade their economic & social status. A relevant example would be Mr Bill Gates (founder of Microsoft Corporation) He had a modest background but he used his talent and motivation to set up the worlds largest technology company.

However, there are some problems that children from poor backgrounds encounter. Many of these children who are 'robbed' of their childhood eg while working, may feel cheated. They often turn to crime. This however, is a small group.

To sum up, children with modest backgrounds are able to deal with problems of adult life because of early exposure, family role models and sheer motivation.

Text ID: 25

I do agree to the statement that children brought up in poor families are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

Offpsring of poor parents are prematurely exposed to obstacles of adult life e.g. learning to survive on a low family income and sacrificing luxuries for essential items. These children began to witness the 'realities' of life in their home or social environment. Their parent's own struggles serve as an example to them.

These children are taught necessary skills for survival as an adult from a very early age. Numerous children eg labor in the weekends to potentially collect some pocket money or even contribute to their household's income. A good example is the many children who accompany their parents to sell produce at the market. They are making a straight contribution to their families in terms of labor or earnings.

Children of poor families also are highly motivated. They tend to set high goals to improve their economic & social situation. A relevant example would be Mr Bill Gates (founder of Microsoft Corporation) He had an impoverished background but he used his expertise and motivation to set up the worlds ultimate computer organization.

Nevertheless, there are some setbacks that children from poor backgrounds encounter. Many of these children who are 'robbed' of their childhood eg while working, may feel cheated. They often turn to crime. This however, is a small group.

In summing up, children with impoverished backgrounds are able to deal with problems of adult life because of early exposure, family role models and sheer motivation.

Text ID: 26

I do agree with the statement that children brought up in poor families are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

Offspring of poor parents are prematurely exposed to obstacles of adult life e.g. learning to survive on a low family income and sacrificing luxuries for essential items. These children began to witness the 'realities' of life in their home or social environment. Their parent's own struggles serve as an example to them.

These children are taught necessary skills for survival as an adult from a very early age. Numerous children eg labor on the weekend to potentially earn some pocket money or even contribute to their household's income. A good example is the many children who accompany their parents to sell produce at the market. They are making a direct contribution to their families in terms of labor or earnings.

Children of poor families also are highly motivated. They tend to set high goals to improve their economic & social situation. A relevant example would be Mr Bill Gates (founder of Microsoft Corporation) He had an impoverished background but he used his expertise and motivation to set up the worlds ultimate technology company.

Nevertheless, there are some setbacks that children from poor backgrounds encounter. Many of these children who are 'robbed' of their childhood eg while working, may feel cheated. They often turn to crime. This however, is a small group.

To sum up, children with impoverished backgrounds are able to deal with problems of adult life because of early exposure, family role models and sheer motivation.

Text ID: 27

I do agree to the statement that children brought up in poor families are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

Children of working-class parents are prematurely exposed to problems of adult life e.g. learning to survive on a low family income and shun luxuries for essential items. These children began to see the 'realities' of life in their home or sociocultural environment. Their parent's own struggles provide an initial impetus to them.

These children are taught requisite skills for survival as an adult from a very early age. Many children eg work in the weekends to either collect some pocket money or even potentially augment their families' income. A quintessential example is the many children who accompany their parents to peddle merchandise at the market. They are making a straight contribution to their families in terms of labor or income.

Children of poor families also are highly motivated. They tend to set high goals to remedy their economic & social situation. A relevant example would be Mr Bill Gates (founder of Microsoft Corporation) He had an impoverished background but he used his talent and motivation to set up the worlds largest computer organization.

However, there are some problems that children from poor backgrounds encounter. Many of these children who are 'robbed' of their childhood eg while working, may feel victimized. They often turn to crime. This however, is a small group.

In summing up, children with impoverished backgrounds are able to deal with problems of adult life because of first-hand experience, family role models and steady resolve.

Text ID: 28

I do agree with the statement that children brought up in poor families are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

Children of working-class parents are prematurely exposed to problems of adult life e.g. learning to survive on a low family income and shun luxuries for essential items. These children began to see the 'realities' of life in their home or sociocultural environment. Their parent's own struggles provide an initial impetus to them.

These children are taught requisite skills for survival as an adult from a very early age. Many children eg work on the weekend to either earn some pocket money or even potentially augment their families' income. A quintessential example is the many children who accompany their parents to peddle merchandise at the market. They are making a direct contribution to their families in terms of labor or income.

Children of poor families also are highly motivated. They tend to set high goals to remedy their economic & social situation. A relevant example would be Mr Bill Gates (founder of Microsoft Corporation) He had an impoverished background but he used his talent and motivation to set up the worlds largest technology company.

However, there are some problems that children from poor backgrounds encounter. Many of these children who are 'robbed' of their childhood eg while working, may feel victimized. They often turn to crime. This however, is a small group.

To sum up, children with impoverished backgrounds are able to deal with problems of adult life because of first-hand experience, family role models and steady resolve.

Text ID: 29

I do agree to the statement that children brought up in poor families are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

Children of poor parents are prematurely exposed to predicaments of adult life e.g. learning to survive on a low family income and sacrificing luxuries for essential items. These children began to glimpse the 'realities' of life in their home or social environment. Their parent's own strife serve as an example to them.

These children are taught necessary skills for survival as an adult from a very early age. Innumerable children eg hustle in the weekends to conceivably collect some pocket money or even contribute to their families' income. A good example is the many children who accompany their parents to sell produce at the market. They are making a straight contribution to their families in terms of labor or subsistence.

Children of poor families also are highly motivated. They tend to set high goals to improve their economic & social situation. A relevant example would be Mr Bill Gates (founder of Microsoft Corporation) He had an impoverished background but he used his ingenuity and perseverance to set up the worlds pre-eminent computer organization.

Notwithstanding, there are some problems that children from poor backgrounds weather. Many of these children who are 'robbed' of their childhood eg while working, may feel cheated. They often turn to crime. This however, is a small group.

In summing up, children with impoverished backgrounds are able to deal with problems of adult life because of early exposure, family role models and sheer motivation.

Text ID: 30

I do agree with the statement that children brought up in poor families are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

Children of poor parents are prematurely exposed to predicaments of adult life e.g. learning to survive on a low family income and sacrificing luxuries for essential items. These children began to glimpse the 'realities' of life in their home or social environment. Their parent's own strife serve as an example to them.

These children are taught necessary skills for survival as an adult from a very early age. Innumerable children eg hustle on the weekend to conceivably earn some pocket money or even contribute to their families' income. A good example is the many children who accompany their parents to sell produce at the market. They are making a direct contribution to their families in terms of labor or subsistence.

Children of poor families also are highly motivated. They tend to set high goals to improve their economic & social situation. A relevant example would be Mr Bill Gates (founder of Microsoft Corporation) He had an impoverished background but he used his ingenuity and perseverance to set up the worlds pre-eminent technology company.

Notwithstanding, there are some problems that children from poor backgrounds weather. Many of these children who are 'robbed' of their childhood eg while working, may feel cheated. They often turn to crime. This however, is a small group.

To sum up, children with impoverished backgrounds are able to deal with problems of adult life because of early exposure, family role models and sheer motivation.

Appendix G IELTS Writing Task 2: Public band descriptors

Band	Task response	Coherence and cohesion	Lexical resource	Grammatical range and accuracy
9	• fully addresses all parts of the task • presents a fully developed position in answer to the question with relevant, fully extended and well supported ideas	• uses cohesion in such a way that it attracts no attention • skilfully manages paragraphing	• uses a wide range of vocabulary with very natural and sophisticated control of lexical features; rare minor errors occur only as 'slips'	• uses a wide range of structures with full flexibility and accuracy; rare minor errors occur only as 'slips'
8	• sufficiently addresses all parts of the task • presents a well-developed response to the question with relevant, extended and supported ideas	• sequences information and ideas logically • manages all aspects of cohesion well • uses paragraphing sufficiently and appropriately	• uses a wide range of vocabulary fluently and flexibly to convey precise meanings • skilfully uses uncommon lexical items but there may be occasional inaccuracies in word choice and collocation • produces rare errors in spelling and/or word formation	• uses a wide range of structures • the majority of sentences are error-free • makes only very occasional errors or inappropriacies
7	• addresses all parts of the task • presents a clear position throughout the response • presents, extends and supports main ideas, but there may be a tendency to over-generalise and/or supporting ideas may lack focus	• logically organises information and ideas; there is clear progression throughout • uses a range of cohesive devices appropriately although there may be some under-/over-use • presents a clear central topic within each paragraph	• uses a sufficient range of vocabulary to allow some flexibility and precision • uses less common lexical items with some awareness of style and collocation • may produce occasional errors in word choice, spelling and/or word formation	• uses a variety of complex structures • produces frequent error-free sentences • has good control of grammar and punctuation but may make a few errors
6	• addresses all parts of the task although some parts may be more fully covered than others • presents a relevant position although the conclusions may become unclear or repetitive • presents relevant main ideas but some may be inadequately developed/unclear	• arranges information and ideas coherently and there is a clear overall progression • uses cohesive devices effectively, but cohesion within and/or between sentences may be faulty or mechanical • may not always use referencing clearly or appropriately • uses paragraphing, but not always logically	• uses an adequate range of vocabulary for the task • attempts to use less common vocabulary but with some inaccuracy • makes some errors in spelling and/or word formation, but they do not impede communication	• uses a mix of simple and complex sentence forms • makes some errors in grammar and punctuation but they rarely reduce communication
5	• addresses the task only partially; the format may be inappropriate in places • expresses a position but the development is not always clear and there may be no conclusions drawn • presents some main ideas but these are limited and not sufficiently developed; there may be irrelevant detail	• presents information with some organisation but there may be a lack of overall progression • makes inadequate, inaccurate or over-use of cohesive devices • may be repetitive because of lack of referencing and substitution • may not write in paragraphs, or paragraphing may be inadequate	• uses a limited range of vocabulary, but this is minimally adequate for the task • may make noticeable errors in spelling and/or word formation that may cause some difficulty for the reader	• uses only a limited range of structures • attempts complex sentences but these tend to be less accurate than simple sentences • may make frequent grammatical errors and punctuation may be faulty; errors can cause some difficulty for the reader
4	• responds to the task only in a minimal way or the answer is tangential; the format may be inappropriate • presents a position but this is unclear • presents some main ideas but these are difficult to identify and may be repetitive, irrelevant or not well supported	• presents information and ideas but these are not arranged coherently and there is no clear progression in the response • uses some basic cohesive devices but these may be inaccurate or repetitive • may not write in paragraphs or their use may be confusing	• uses only basic vocabulary which may be used repetitively or which may be inappropriate for the task • has limited control of word formation and/or spelling; errors may cause strain for the reader	• uses only a very limited range of structures with only rare use of subordinate clauses • some structures are accurate but errors predominate, and punctuation is often faulty
3	• does not adequately address any part of the task • does not express a clear position • presents few ideas, which are largely undeveloped or irrelevant	• does not organise ideas logically • may use a very limited range of cohesive devices, and those used may not indicate a logical relationship between ideas	• uses only a very limited range of words and expressions with very limited control of word formation and/or spelling • errors may severely distort the message	• attempts sentence forms but errors in grammar and punctuation predominate and distort the meaning
2	• barely responds to the task • does not express a position • may attempt to present one or two ideas but there is no development	• has very little control of organisational features	• uses an extremely limited range of vocabulary; essentially no control of word formation and/or spelling	• cannot use sentence forms except in memorised phrases
1	• answer is completely unrelated to the task	• fails to communicate any message	• can only use a few isolated words	• cannot use sentence forms at all
0	• does not attend • does not attempt the task in any way • writes a totally memorised response			

Figure 12 IELTS Writing Task 2: Public band descriptors (IELTS, n.d.-c)

Appendix H Holistic writing assessment scale

Score	Skill level	Description
9	Expert user	The test taker has fully operational command of the language. Their use of English is appropriate, accurate and fluent, and shows complete understanding.
8	Very good user	The test taker has fully operational command of the language with only occasional unsystematic inaccuracies and inappropriate usage. They handle complex and detailed argumentation well.
7	Good user	The test taker has operational command of the language, though with occasional inaccuracies and inappropriate usage in some situations. They generally handle complex language well.
6	Competent user	The test taker has an effective command of the language despite some inaccuracies and inappropriate usage. They can use fairly complex language.
5	Modest user	The test taker has a partial command of the language and copes with overall meaning, although they are likely to make many mistakes. They are able to handle basic communication.
4	Limited user	The test taker's basic competence is limited. They frequently show problems in expression. They are not able to use complex language.
3	Extremely limited user	The test taker conveys and understands only general meaning. There are frequent breakdowns in communication.
2	Intermittent user	The test taker has great difficulty using any written English.
1	Non-user	The test taker has no ability to use the language except a few isolated words.

Figure 13 Holistic writing assessment scale (adapted from IELTS, n.d.-b)

Bibliography

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. (2007). Lexical errors as writing quality predictors. *Studia linguistica, 61*(1), 1-19. https://doi.org/10.1111/j.1467-9582.2007.00127.x

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

AlHassan, L., & Wood, D. (2015). The effectiveness of focused instruction of formulaic sequences in augmenting L2 learners' academic writing skills: A quantitative research study. *Journal of English for Academic Purposes, 17*, 51-62. https://doi.org/10.1016/j.jeap.2015.02.001

Allison, P. (2012, September 10). When can you safely ignore multicollinearity? *Statistical Horizons*. https://statisticalhorizons.com/multicollinearity

Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics, 22*(2), 173-195. https://doi.org/10.1093/applin/22.2.173

Amidei, J., Piwek, P., & Willis, A. (2019). The use of rating and Likert scales in Natural Language Generation human evaluation tasks: A review and some recommendations. *Proceedings of the 12th International Conference on Natural Language Generation*, 397-402. https://doi.org/10.18653/v1/W19-8648

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *International Reading Association* (pp. 77-117).

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., & Liu, S. (2015). Predicting EFL writing ability from levels of mental representation measured by Coh-Metrix: A structural equation modeling study. *Assessing Writing, 24*, 35-58. https://doi.org/10.1016/j.asw.2015.03.001

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

Baayen, H. (2008). *Analyzing linguistic data: a practical introduction to statistics using R*. Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: designing and developing useful language tests*. Oxford University Press.

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

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

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography*, *6*(4), 253-279. https://doi.org/10.1093/ijl/6.4.253

Bellhouse, G. L. (2018). Are teachers given sufficient tools as examiners in high-stakes language testing? A Study of the new foreign language speaking component of the French baccalauréat. In D. Xerri & P. Vella Briffa (Eds.), *Teacher involvement in high-stakes language testing* (pp. 85-103). Springer International Publishing. https://doi.org/10.1007/978-3-319-77177-9_6

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

Biber, D., & Gray, B. (2013a). Discourse characterisitics of writing and speaking task types on the TOEFL IBT® test: a lexico-grammatical analysis. *ETS Research Report Series*, *2013*(1), i-128. https://doi.org/10.1002/j.2333-8504.2013.tb02311.x

Biber, D., & Gray, B. (2013b). Nominalizing the verb phrase in academic science writing. In B. Aarts, G. Leech, J. Close, & S. Wallis (Eds.), *The verb phrase in English: Investigating recent language change with corpora* (pp. 99-132). Cambridge University Press. https://doi.org/10.1017/CBO9781139060998.006

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc.

Boers, F. (2020). Factors affecting the learning of multiword items. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 143-157). Routledge. https://doi.org/10.4324/9780429291586-10

Boers, F., Demecheleer, M., He, L., Deconinck, J., Stengers, H., & Eyckmans, J. (2017). Typographic enhancement of multiword units in second language text. *International Journal of Applied Linguistics, 27*(2), 448-469. https://doi.org/10.1111/ijal.12141

Bonk, W. J. (2000). Second language lexical knowledge and listening comprehension. *International Journal of Listening, 14*(1), 14-31. https://doi.org/10.1080/10904018.2000.10499033

Borg, S. (2003). Teacher cognition in language teaching: A review of research on what language teachers think, know, believe, and do. *Language Teaching, 36*(2), 81-109. https://doi.org/10.1017/S0261444803001903

Borg, S. (2006). *Teacher cognition and language education: Research and practice*. Continuum.

Boyd, E., & Donnarumma, D. (2018). Assessment literacy for teachers: a pilot study investigating the challenges, benefits and impact of assessment literacy training. In D. Xerri & P. Vella Briffa (Eds.), *Teacher involvement in high-stakes language testing* (pp. 105-126). Springer International Publishing. https://doi.org/10.1007/978-3-319-77177-9_7

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics, 36*(1), 1-22. https://doi.org/10.1093/applin/amt018

British Council. (n.d.). *Collocation*. Retrieved Aug 11, 2021 from https://www.teachingenglish.org.uk/article/collocation

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.

Brown, J. D. (1988). *Understanding research in second language learning: a teacher's guide to statistics and research design*. Cambridge University Press.

Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL Quarterly, 25*(4), 587-603. https://doi.org/10.2307/3587078

Browne, C., Culligan, B., & Phillips, J. (2013). *The New General Service List*. Retrieved July 1, 2021 from http://www.newgeneralservicelist.org

Browne, K. (2005). Snowball sampling: Using social networks to research non-heterosexual women. *International Journal of Social Research Methodology: Theory & Practice, 8*(1), 47-60. https://doi.org/10.1080/1364557032000081663

Bryman, A. (2006). Integrating quantitative and qualitative research: how is it done? *Qualitative Research, 6*(1), 97-113. https://doi.org/10.1177/1468794106058877

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of second language writing, 26*, 42-65. https://doi.org/https://doi.org/10.1016/j.jslw.2014.09.005

Bulté, B., Housen, A., Pierrard, M., & Van Daele, S. (2008). Investigating lexical proficiency development over time – the case of Dutch-speaking learners of French in Brussels. *Journal of French Language Studies, 18*(3), 277-298. https://doi.org/10.1017/S0959269508003451

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, 711-733.

Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge University Press. https://doi.org/10.1017/CBO9780511750526

Chan, S., Bax, S., & Weir, C. (2017). Researching participants taking IELTS Academic Writing Task 2 (AWT2) in paper mode and in computer mode in terms of score equivalence, cognitive validity and other factors. *IELTS Research Reports Online Series, 4*, 2-47. https://www.ielts.org/teaching-and-research/research-reports

Charles, M., & Pecorari, D. E. (2016). *Introducing English for Academic Purposes*. Routledge.

Chen, C., & Truscott, J. (2010). The effects of repetition and L1 lexicalization on incidental vocabulary acquisition. *Applied Linguistics, 31*(5), 693-713. https://doi.org/10.1093/applin/amq031

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

Chen, Y.-H., & Baker, P. (2016). Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics, 37*(6), 849-880. https://doi.org/10.1093/applin/amu065

Choi, S. (2017). Processing and learning of enhanced English collocations: An eye movement study. *Language Teaching Research, 21*(3), 403-426. https://doi.org/10.1177/1362168816653271

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.

Clenton, J., & Booth, P. (2020). Introduction: Vocabulary and the four skills – current issues and future concerns. In J. Clenton & P. Booth (Eds.), *Vocabulary and the Four Skills: Pedagogy, Practice, and Implications for Teaching Vocabulary* (pp. 3-19). Routledge.

Cobb, T. *Web Vocabprofile, an adaptation of Heatley, Nation & Coxhead's (2002) Range.* In http://www.lextutor.ca/vp/

Cobb, T. (n.d.). *Compleat Lexical Tutor v.8.3*. Retrieved July 1, 2021 from https://www.lextutor.ca

Cobb, T., & Horst, M. (2015). Learner corpora and lexis. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 185-206). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414

Cobb, T., & Laufer, B. (2021). The Nuclear Word Family List: A list of the most frequent family members, including base and affixed words. *Language Learning, Advance online publication*. https://doi.org/10.1111/lang.12452

Cohen, N., & Arieli, T. (2011). Field research in conflict environments: Methodological challenges and snowball sampling. *Journal of peace research, 48*(4), 423-435. https://doi.org/10.1177/0022343311405698

Coleman, D., Starfield, S., & Hagan, A. (2003). The attitudes of IELTS stakeholders: Student and staff perceptions of IELTS in Australian, UK and Chinese tertiary institutions. In *IELTS Research Reports 2003* (Vol. 5, pp. 160-235). IELTS Australia.

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

Connor-Linton, J. (1995). Crosscultural comparison of writing standards: American ESL and Japanese EFL. *World Englishes, 14*(1), 99-115. https://doi.org/10.1111/j.1467-971X.1995.tb00343.x

Coulmas, F. (1979). On the sociolinguistic relevance of routine formulae. *Journal of Pragmatics, 3*(3), 239-266. https://doi.org/10.1016/0378-2166(79)90033-X

Council of Europe. (2001). *Common European framework of reference for languages: learning, teaching, assessment*. Press Syndicate of the University of Cambridge.

Council of Europe. (2021). *Global scale - Table 1 (CEFR 3.3): Common Reference levels*. Retrieved October 1, 2021 from https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale

Cowie, A., P., & Howarth, P. (1996). Phraseological competence and written proficiency. In G. M. Blue & R. Mitchell (Eds.), *Language and education (British studies in applied linguistics II)* (pp. 80-93). Multilingual Matters.

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., Salsbury, T., & McNamara, D. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning, 59*(2), 307-334. https://doi.org/10.1111/j.1467-9922.2009.00508.x

Crossley, S. A., Kyle, K., Allen, L., Guo, L., & McNamara, D. (2014). Linguistic microfeatures to predict L2 writing proficiency: a case study in automated writing evaluation. *The Journal of Writing Assessment, 7*(1).

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

Crossley, S. A., Salsbury, T., Mcnamara, D. S., & Jarvis, S. (2011). What is lexical proficiency? Some snswers from computational models of speech data. *TESOL Quarterly, 45*(1), 182-193. https://doi.org/10.5054/tq.2010.244019

Crossley, S. A., & Skalicky, S. (2019). Examining lexical development in second language learners: An approximate replication of Salsbury, Crossley & McNamara (2011). *Language Teaching, 52*(3), 385-405. https://doi.org/10.1017/S0261444817000362

Crossley, S. A., Skalicky, S., Kyle, K., & Monteiro, K. (2019). Absolute frequency effects in second language acquisition. *Studies in Second Language Acquisition, 41*(4), 721-744. https://doi.org/10.1017/S0272263118000268

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

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: a descriptive framework. *The Modern Language Journal, 86*(1), 67-96. https://doi.org/10.1111/1540-4781.00137

Cummins, J. (2003). BICS and CALP: Origins and rationale for the distinction. In C. B. Paulston & G. R. Tucker (Eds.), *Sociolinguistics: the essential readings* (pp. 322-328). Wiley.

Dabbagh, A., & Janebi Enayat, M. (2019). The role of vocabulary breadth and depth in predicting second language descriptive writing performance. *The language learning Journal, 47*(5), 575-590. https://doi.org/10.1080/09571736.2017.1335765

Daller, H., Milton, J., & Treffers-Daller, J. (2007). Editors' introduction: Conventions, terminology and an overview of the book. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. 1-32). Cambridge University Press. https://doi.org/10.1017/CBO9780511667268.003

Daller, H., & Phelan, D. (2007). What is in a teacher's mind? Teacher ratings of EFL essays and different aspects of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. 234-244). Cambridge University Press. https://doi.org/10.1017/CBO9780511667268.016

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

Daller, H., & Xue, H. (2007). Lexical richness and the oral proficiency of Chinese EFL students. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. 150-164). Cambridge University Press. https://doi.org/10.1017/CBO9780511667268.011

Daller, M., Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics, 24*(2), 197-222. https://doi.org/10.1093/applin/24.2.197

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

Dang, T. N. Y. (2020). Corpus-based word lists in second language vocabulary research, learning, and teaching. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 288-303). Routledge. https://doi.org/10.4324/9780429291586-19

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

Davis, K. A. (1995). Qualitative theory and methods in applied linguistics research. *TESOL Quarterly, 29*(3), 427-453. https://doi.org/10.2307/3588070

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language testing, 33*(1), 117-135. https://doi.org/10.1177/0265532215582282

De Cock, S., Granger, S., Leech, G., & McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learner English on computer* (pp. 67-79). Longman.

de Groot, A. M. B., & Poot, R. (1997). Word translation at three levels of proficiency in a second language: the ubiquitous Involvement of conceptual memory. *Language Learning, 47*(2), 215-264. https://doi.org/10.1111/0023-8333.71997007

DeVellis, R. F. (2003). *Scale development: theory and applications* (2nd ed.). SAGE Publications.

Dóczi, B., & Kormos, J. (2016). *Longitudinal developments in vocabulary knowledge and lexical organization*. Oxford University Press.

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. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (Section H)*. Council of Europe/Language Policy Division.

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

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement: Analyzing and evaluating rater-mediated assessments* (2nd, Ed.). Peter Lang.

Eckes, T., Müller-Karabil, A., & Zimmermann, S. (2016). Assessing writing. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (Vol. 12, pp. 147-164). De Gruyter. https://doi.org/10.1515/9781614513827-012

Elgort, I., & Siyanova-Chanturia, A. (2021). Interdisciplinary approaches to researching L2 lexical acquisition, processing, and use: An introduction to the special issue. *Second Language Research*. https://doi.org/10.1177/0267658320988050

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition*, *18*(1), 91-126. http://www.jstor.org/stable/44487860

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*(2), 143-188. https://doi.org/10.1017/S0272263102002024

Ellis, N. C. (2004). The processes of second language acquisition. In B. VanPatten, J. Williams, S. Rott, & M. Overstreet (Eds.), *Form-meaning connections in second language acquisition* (pp. 51-80). Routledge. https://doi.org/10.4324/9781410610607

Ellis, N. C., & Simpson-Vlach, R. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus linguistics and linguistic theory, 5*(1), 61-78. https://doi.org/10.1515/CLLT.2009.003

Ellis, N. C., & Wulff, S. (2015). Usage-based approaches to SLA. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: an introduction* (2nd ed., pp. 75-93). Routledge.

Ellis, R. (1995). Modified oral input and the acquisition of word meanings. *Applied Linguistics, 16*(4), 409-441. https://doi.org/10.1093/applin/16.4.409

Ellis, R. (2008). *The study of second language acquisition* (2nd ed.). Oxford University Press.

Ellis, R., Tanaka, Y., & Yamazaki, A. (1994). Classroom interaction, comprehension, and the acquisition of L2 word meanings. *Language Learning, 44*(3), 449-491. https://doi.org/10.1111/j.1467-1770.1994.tb01114.x

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition, 26*(1), 59-84. https://doi.org/10.1017/S0272263104026130

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of second language writing, 4*(2), 139-155. https://doi.org/10.1016/1060-3743(95)90004-7

EnglishProfile. (2015). *Text Inspector*. Cambridge University Press. http://www.englishprofile.org/wordlists/text-inspector

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater ® scoring. *Language testing, 27*(3), 317-334. https://doi.org/10.1177/0265532210363144

Erguvan, I. D., & Aksu Dunya, B. (2020). Analyzing rater severity in a freshman composition course using many facet Rasch measurement. *Language Testing in Asia, 10*(1), 1. https://doi.org/10.1186/s40468-020-0098-3

Estaji, M., & Ghiasvand, F. (2019). The washback effect of IELTS examination on EFL teachers' perceived sense of professional identity: Does IELTS related experience make a difference? *Journal of Modern Research in English Language Studies, 6*(3), 103-183. https://doi.org/10.30479/jmrels.2019.11123.1391

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

Fan, M. (2009). An exploratory study of collocational use by ESL students – A task based approach. *System, 37*(1), 110-123. https://doi.org/10.1016/j.system.2008.06.004

Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly, 28*(2), 414-420. https://doi.org/10.2307/3587446

Firth, J. R. (1957). A synopsis of linguistic theory 1930–55. In *Studies in linguistic analysis* (pp. 1-32). The Philological Society.

Foster, P., & Tavakoli, P. (2009). Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity. *Language Learning, 59*(4), 866-896. https://doi.org/10.1111/j.1467-9922.2009.00528.x

Friedline, B. E. (2011). *Challenges in the second language acquisition of derivational morphology: from theory to practice* [Doctoral Dissertation, University of Pittsburgh]. http://d-scholarship.pitt.edu/8351/

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

Gao, X., & Ma, Q. (2011). Vocabulary learning and teaching beliefs of pre-service and in-service teachers in Hong Kong and mainland China. *Language Awareness, 20*(4), 327-342. https://doi.org/10.1080/09658416.2011.579977

Gardner, D., & Davies, M. (2014). A new Academic Vocabulary List. *Applied Linguistics, 35*(3), 305-327. https://doi.org/10.1093/applin/amt015

Garner, J., & Crossley, S. (2018). A latent curve model approach to studying L2 n-gram development. *The Modern Language Journal, 102*(3), 494-511. https://doi.org/10.1111/modl.12494

Gilquin, G. (2007). To err is not at all: What corpus and elicitation can reveal about the use of collocations by learners. *Zeitschrift für Anglistik und Amerikanistik, 55*(3), 273-291. https://doi.org/10.1515/zaa.2007.55.3.273

Ginther, A., & Grant, L. (1997). The influences of proficiency, language background and topic on the production of grammatical form and error in the Test of Written English. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current Developments and Alternatives in Language Assessment: Proceedings of LTRC 96* (pp. 385-397). Universities of Tempere and Jyväskylä.

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

Goldberg, A. E. (2006). *Constructions at work the nature of generalization in language*. Oxford University Press.

Goldberg, A. E. (2013). Constructionist approaches. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford handbook of construction grammar* (pp. 15-31). Oxford University Press.

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

González-Fernández, B., & Schmitt, N. (2019). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics, 41*(4), 481-505. https://doi.org/10.1093/applin/amy057

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments & Computers, 36*(2), 193-202. https://doi.org/10.3758/BF03195564

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. https://www.jstor.org/stable/24157525

Granger, S. (2019). Formulaic sequences in learner corpora: Collocations and lexical bundles. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding Formulaic Language: A Second Language Acquisition Perspective* (pp. 228-247). Routledge.

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

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of second language writing, 9*(2), 123-145. https://doi.org/10.1016/S1060-3743(00)00019-9

Green, A. (2019). Restoring perspective on the IELTS test. *ELT Journal, 73*(2), 207-215. https://doi.org/10.1093/elt/ccz008

Gries, S. T., & Wulff, S. (2005). Do foreign language learners also have constructions? *Annual review of cognitive linguistics, 3*(1), 182-200. https://doi.org/10.1075/arcl.3.10gri

Griffin, P. (1990). *Characteristics of the Test Components of the IELTS Battery: Australian trial data* RELC annual seminar, Singapore.

Grobe, C. (1981). Syntactic maturity, mechanics, and vocabulary as predictors of quality ratings. *Research in the Teaching of English, 15*(1), 75-85. http://www.jstor.org/stable/40170871

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. (2013). Looking at L2 vocabulary knowledge dimensions from an assessment perspective – challenges and potential solutions. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *EUROSLA Monographs Series 2* (pp. 11-28). European Second Language Association.

Gyllstad, H., & Schmitt, N. (2019). Testing formulaic language. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding Formulaic Language: A Second Language Acquisition Perspective* (pp. 174-191). Routledge. https://doi.org/10.4324/9781315206615

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).

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2014). *Halliday's Introduction to Functional Grammar*. Routledge. https://doi.org/10.4324/9780203431269

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.

Harrington, M. (2018). Lexical facility and IELTS performance. In *Lexical Facility: Size, Recognition Speed and Consistency as Dimensions of Second Language Vocabulary Knowledge* (pp. 187-203). Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-37262-8_8

Hashemi, M. R. (2012). Reflections on mixing methods in applied linguistics research. *Applied Linguistics, 33*(2), 206-212. https://doi.org/10.1093/applin/ams008

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

Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing, 9*(2), 122-159. https://doi.org/10.1016/j.asw.2004.06.001

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). *RANGE and FREQUENCY programs.* In http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

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.

Henriksen, B., & Stæhr, L. S. (2009). Processes in the development of L2 collocational knowledge — a challenge for language learners, researchers and teachers. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: multiple interpretations* (pp. 224-231). Palgrave Macmillan.

Hill, J. (2000). Revising priorities: From grammatical failure to collocational success. In M. Lewis (Ed.), *Teaching collocation: further developments in the lexical approach* (pp. 47-70). Language Teaching Publications.

Hill, K., Storch, N., & Lynch, B. (1999). A comparison of IELTS and TOEFL as predictors of academic success. *IELTS Research Reports, 2*, 62-73.

Holzknecht, F., Kremmel, B., Konzett, C., Eberharter, K., & Spöttl, C. (2018). Potentials and challenges of teacher involvement in rating scale design for high-stakes exams. In D. Xerri & P. Vella Briffa (Eds.), *Teacher Involvement in High-Stakes Language Testing* (pp. 47-66). Springer International Publishing. https://doi.org/10.1007/978-3-319-77177-9_4

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. https://spacy.io/

Horst, M., & Collins, L. (2006). From faible to strong: How does their vocabulary grow? *Canadian Modern Language Review, 63*(1), 83-106. https://doi.org/10.3138/cmlr.63.1.83

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

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a foreign language, 13*, 403-430.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press.

Hunston, S., & Francis, G. (2000). *Pattern Grammar: A corpus-driven approach to the lexical grammar of English*. John Benjamins.

Hyland, K. (2012). Bundles in academic discourse. *Annual Review of Applied Linguistics, 32*, 150-169. https://doi.org/10.1017/S0267190512000037

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269-293). Penguin Education.

IELTS. (2018). *Test tasker performance 2018*. Retrieved December 1, 2020 from https://www.ielts.org/en-us/research/test-taker-performance

IELTS. (2019). *IELTS grows to 3.5 million a year*. Retrieved December 1, 2020 from https://www.ielts.org/en-us/news/2019/ielts-grows-to-three-and-a-half-million-a-year

IELTS. (n.d.-a). *Comparing IELTS and the Common European Framework*. Retrieved December 1, 2020 from https://www.ielts.org/-/media/pdfs/comparing-ielts-and-cefr.ashx

IELTS. (n.d.-b). *How IELTS is scored*. Retrieved December 1, 2020 from https://www.ielts.org/about-the-test/how-ielts-is-scored

IELTS. (n.d.-c). *IELTS Scoring in Detail*. Retrieved October 15, 2021 from https://www.ielts.org/en-us/ielts-for-organisations/ielts-scoring-in-detail

IELTS. (n.d.-d). *Sample test questions*. Retrieved December 1, 2020 from https://www.ielts.org/en-us/about-the-test/sample-test-questions

IELTS. (n.d.-e). *Test format*. Retrieved December 1, 2020 from https://www.ielts.org/en-us/about-the-test/test-format

Jackendoff, R. S. (2002). What's in the lexicon? In S. Nooteboom, F. Weerman, & F. Wijnen (Eds.), *Storage and computation in the language faculty* (Vol. 30, pp. 23-58). Springer. https://doi.org/10.1007/978-94-010-0355-1_2

James, C. (1998). *Errors in language learning and use: exploring error analysis*. Longman.

Jarvis, S. (2013a). Capturing the diversity in lexical diversity. *Language Learning, 63*(1), 87-106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

Jarvis, S. (2013b). 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. (2021). Phraseological complexity and low- and intermediate-level L2 learners' writing quality. *International Review of Applied Linguistics in Language Teaching*. https://doi.org/10.1515/iral-2019-0147

Jiang, N., & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *The Modern Language Journal, 91*(3), 433-445. https://doi.org/10.1111/j.1540-4781.2007.00589.x

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language testing, 26*(4), 485-505. https://doi.org/10.1177/0265532209340186

Jones, M., & Haywood, S. (2004). Facilitating the acquisition of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 269-300). John Benjamins.

Juffs, A. (2009). Second language acquisition of the lexicon. In W. Ritchie & T. K. Bhatia (Eds.), *The New Handbook of Second Language Acquisition* (pp. 181-205). Emerald.

Juffs, A. (2019). Lexical development in the writing of intensive English program students. In R. M. DeKeyser & G. Prieto Botana (Eds.), *Doing SLA research with implications for the classroom: Reconciling methodological demands and pedagogical applicability* (pp. 179-220). John Benjamins.

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

Kang, B.-m. (2018). Collocation and word association: Comparing collocation measuring methods. *International Journal of Corpus Linguistics, 23*(1), 85-113. https://doi.org/10.1075/ijcl.15116.kan

Kellogg, R. T., & Raulerson, B. A. (2007). Improving the writing skills of college students. *Psychonomic bulletin & review, 14*(2), 237-242. https://doi.org/10.3758/BF03194058

Khalil, A. (1985). Communicative error evaluation: Native speakers' evaluation and interpretation of written errors of Arab EFL learners. *TESOL Quarterly, 19*(2), 335-351. https://doi.org/10.2307/3586833

Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal, 102*(1), 120-141. https://doi.org/10.1111/modl.12447

Koda, K. (1993). Task-induced variability in FL composition: Language-specific perspectives. *Foreign Language Annals, 26*(3), 332-346. https://doi.org/10.1111/j.1944-9720.1993.tb02290.x

Koizumi, R., & In'nami, Y. (2020). Structural equation modeling of vocabulary size and depth using conventional and bayesian methods. *Frontiers in Psychology, 11*(618). https://doi.org/10.3389/fpsyg.2020.00618

Kormos, J. (2006). *Speech production and second language acquisition*. Lawrence Erlbaum Associates Publishers.

Kroll, J., & De Groot, A. M. B. (Eds.). (1997). *Tutorials in Bilingualism: psycholinguistic perspectives*. Erlbaum.

Kroll, J. F., & Merves, J. S. (1986). Lexical access for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*(1), 92-107. https://doi.org/10.1037/0278-7393.12.1.92

Kyle, K. (2020). Measuring lexical richness. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 454-476). Routledge. https://doi.org/10.4324/9780429291586-29

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. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of second language writing, 34*, 12-24. https://doi.org/10.1016/j.jslw.2016.10.003

Kyle, K., & Crossley, S. A. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language testing, 34*(4), 513-535. https://doi.org/10.1177/0265532217712554

Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior Research Methods, 50*(3), 1030-1046. https://doi.org/10.3758/s13428-017-0924-4

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

Lado, R. (1961). *The construction and use of foreign language tests*. McGraw-Hill.

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. (1992). Reading in a foreign language: how does L2 lexical knowledge interact with the reader's general academic ability. *Journal of Research in Reading, 15*(2), 95-103. https://doi.org/10.1111/j.1467-9817.1992.tb00025.x

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: do we need both to measure vocabulary knowledge? *Language testing, 21*(2), 202-226. https://doi.org/10.1191/0265532204lt277oa

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning, 54*(3), 399-436. https://doi.org/10.1111/j.0023-8333.2004.00260.x

Laufer, B., & Nation, I. S. P. (2012). Vocabulary. In S. M. Gass & A. Mackey (Eds.), *The Routledge Handbook of Second Language Acquisition* (pp. 163-176). Taylor & Francis.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307-322. https://doi.org/10.1093/applin/16.3.307

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*. https://doi.org/doi:10.1515/iral-2018-0220

Lee, S. H. (2003). ESL learners' vocabulary use in writing and the effects of explicit vocabulary instruction. *System, 31*(4), 537-561. https://doi.org/10.1016/j.system.2003.02.004

Leki, I., & Carson, J. G. (1994). Students' perceptions of EAP writing instruction and writing needs across the disciplines. *TESOL Quarterly, 28*(1), 81-101. https://doi.org/10.2307/3587199

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

Levelt, W. J. M. (1989). *Speaking: from intention to articulation*. MIT Press.

Levelt, W. J. M. (2001). Spoken word production: A theory of lexical access. *Proceedings of the National Academy of Sciences, 98*(23), 13464-13471. https://doi.org/10.1073/pnas.231459498

Levitzky-Aviad, T., & Laufer, B. (2013). Lexical properties in the writing of foreign language learners over eight years of study: Single words and collocations. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *EUROSLA Monographs Series 2* (pp. 127-148). European Second Language Association.

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

Lewis, M. (1993). *The lexical approach: the state of ELT and a way forward*. Language Teaching Publications.

Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: a longitudinal case study. *Journal of second language writing, 18*(2), 85-102. https://doi.org/10.1016/j.jslw.2009.02.001

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

Linacre, J. M. (2021). *Winsteps*®. In (Version 5.1.4) [Computer software]. Winsteps.com. https://www.winsteps.com

Lindqvist, C., Bardel, C., & Gudmundson, A. (2011). Lexical richness in the advanced learner's oral production of French and Italian L2. *49*(3), 221-240. https://doi.org/10.1515/iral.2011.013

Lindqvist, C., Gudmundson, A., & Bardel, C. (2013). A new approach to measuring lexical sophistication in L2 oral production. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *EUROSLA Monographs Series 2* (pp. 109-126). European Second Language Association.

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

Lorenz, G. R. (1999). *Adjective intensification – Learner's versus native speakers: a corpus study of argumentative Writing*. Brill.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474-496. https://doi.org/10.1075/ijcl.15.4.02lu

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly, 45*(1), 36-62. http://www.jstor.org/stable/41307615

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal, 96*(2), 190-208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

Lumley, T. (1998). Perceptions of language-trained raters and occupational experts in a test of occupational English language proficiency. *English for Specific Purposes, 17*(4), 347-367. https://doi.org/10.1016/S0889-4906(97)00016-1

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/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

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Transcription format and programs, Vol. 1, 3rd ed.* Lawrence Erlbaum Associates Publishers.

Malvern, D. D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: quantification and assessment*. Palgrave Macmillan. https://doi.org/978-1-4039-0231-3

Malvern, D. D., & Richards, B. J. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (Eds.), *Evolving models of language* (pp. 58-71). Multilingual Matters.

Martin, K. I., & Tokowicz, N. (2020). The grammatical class effect is separable from the concreteness effect in language learning. *Bilingualism, 23*(3), 554-569. https://doi.org/10.1017/S1366728919000233

McCarthy, P. M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)* (Publication Number 3199485) [Doctoral dissertation, University of Memphis]. ProQuest Dissertations & Theses Global.

McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language testing, 24*(4), 459-488. https://doi.org/10.1177/0265532207080767

McIntosh, C., Francis, B., & Poole, R. (Eds.). (2009). *Oxford Collocations Dictionary: For Students of English* (2nd ed.). Oxford University Press.

McNamara, T. F. (1991). Test dimensionality: IRT analysis of an ESP listening test1. *Language testing, 8*(2), 139-159. https://doi.org/10.1177/026553229100800204

McNamara, T. F. (1996). *Measuring second language performance*. Longman.

McNamara, T. F., & Adams, R. J. (1991/1994). Exploring Rater Behaviour with Rasch Techniques. In *Selected papers of the 13th Language Testing Research Colloquium (LTRC).* (pp. 1-29). Educational Testing Service.

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

Meara, P., & Bell, H. (2001). P-Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect, 16*(3), 5-19.

Mickan, P. (2003). 'What's your score?': An investigation into language descriptors for rating written performance. *IELTS Research Reports, 5*, 125-155.

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

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *EUROSLA Monographs Series 2* (pp. 57-78). European Second Language Association.

Milton, J., & Alexiou, T. (2009). Vocabulary size and the Common European Framework of Reference for languages. In B. Richards, M. H. Daller, D. D. Malvern, P. Meara, J. Milton, & J. Treffers-Daller (Eds.), *Vocabulary studies in first and second language acquisition: the interface between theory and application* (pp. 194-211). Palgrave Macmillan UK. https://doi.org/10.1057/9780230242258_12

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in English as a foreign language. In D. R. Chacón-Beltrán, C. Abello-Contesse, & M. d. M. Torreblanca-López (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 83-98). Channel View Publications. https://doi.org/10.21832/9781847692900-007

Mollin, S. (2009). Combining corpus linguistic and psychological data on word co-occurrences: Corpus collocates versus word associations. *Corpus linguistics and linguistic theory, 5*(2), 175-200. https://doi.org/10.1515/CLLT.2009.008

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

Morgado, F. F. R., Meireles, J. F. F., Neves, C. M., Amaral, A. C. S., & Ferreira, M. E. C. (2017). Scale development: ten main limitations and recommendations to improve future research practices. *Psicologia: Reflexão e Crítica, 30*(3), 1-20. https://doi.org/10.1186/s41155-016-0057-1

Morris, L., & Cobb, T. (2004). Vocabulary profiles as predictors of the academic performance of Teaching English as a Second Language trainees. *System, 32*(1), 75-87. https://doi.org/10.1016/j.system.2003.05.001

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., Han, N.-R., Juffs, A., Hill, B., & Zheng, D. (2018). Accurate measurement of lexical sophistication with reference to ESL learner data. *Proceedings of the 11th International*

Conference on Educational Data Mining, Buffalo, New York, 259-265. https://educationaldatamining.org/EDM2018/proceedings/

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

Naismith, B., Juffs, A., & Han, N.-R. (forthcoming). Handle it in house: Learner corpora frequency lists and lexical sophistication. *International Journal of Corpus Linguistics*.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Newbury House Publishers.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*(1), 59-82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P. (2007). Fundamental issues in modelling and assessing vocabulary knowledge. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 35-44). Cambridge University Press. https://doi.org/10.1017/CBO9780511667268.004

Nation, I. S. P. (2011). Research into practice: Vocabulary. *Language Teaching, 44*(4), 529-539. https://doi.org/10.1017/S0261444811000267

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

Nation, I. S. P. (2014). How much input do you need to learn the most frequent 9,000 words? *Reading in a foreign language, 26*, 1-16.

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. John Benjamins.

Nation, I. S. P. (2020). The different aspects of vocabulary knowledge. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 15-29). Routledge. https://doi.org/10.4324/9780429291586-2

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

Nation, I. S. P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: description, acquisition and pedagogy* (pp. 6-19). Cambridge University Press.

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.

O'Brien, R. M. (2017). Dropping highly collinear variables from a model: Why it typically is not a good idea*. *Social Science Quarterly, 98*(1), 360-375. https://doi.org/10.1111/ssqu.12273

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/https://doi.org/10.1111/lang.12427

Oliver, R., Vanderford, S., & Grote, E. (2012). Evidence of English language proficiency and academic achievement of non-English-speaking background students. *Higher Education Research & Development, 31*(4), 541-555. https://doi.org/10.1080/07294360.2011.653958

Open Doors. (2021). *Fields of Study*. Retrieved December 8, 2021 from https://opendoorsdata.org/data/international-students/fields-of-study/

Orr, M. (2002). The FCE Speaking test: using rater reports to help interpret test scores. *System, 30*(2), 143-154. https://doi.org/10.1016/S0346-251X(02)00002-7

Palmer, H. (1933). Aids to conversational skill. *The Bulletin of the Institute for Research in English Teaching, 90*, 1-3.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research, 35*(1), 121-145. https://doi.org/10.1177/0267658317694221

Parent, K. (2019). Do language learners walk the walk? Noun-verb converted forms in learner English. *언어학 연구 (Linguistic Studies), 50*, 215-244.

Parr, J. M., & Timperley, H. S. (2010). Feedback to writing, assessment for teaching and learning and student progress. *Assessing Writing, 15*(2), 68-85. https://doi.org/10.1016/j.asw.2010.05.004

Pearson, W. S. (2019). Critical perspectives on the IELTS test. *ELT Journal, 73*(2), 197-206. https://doi.org/10.1093/elt/ccz006

Pellicer-Sánchez, A. (2017). Learning L2 collocations incidentally from reading. *Language Teaching Research, 21*(3), 381-402. https://doi.org/10.1177/1362168815618428

Pellicer-Sánchez, A. (2020). Learning single words vs. multiword items. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 158-173). Routledge. https://doi.org/10.4324/9780429291586-11

Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro, P. Reitsma, & L. Verhoven (Eds.), *Precursors of Functional Literacy* (pp. 189-213). John Benjamins.

Picoral, A., & Carvalho, A. M. (2020). The acquisition of preposition + article contractions in L3 Portuguese among different L1-speaking learners: a variationist approach. *Languages, 5*(45), 1-17. https://doi.org/10.3390/languages5040045

Pittsburgh. (2021, December 10). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Pittsburgh&oldid=1059267174

Polio, C., & Glew, M. (1996). ESL writing assessment prompts: How students choose. *Journal of second language writing, 5*(1), 35-49. https://doi.org/10.1016/S1060-3743(96)90014-4

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

Pula, J. J., & Huot, B. A. (1993). A model of background influences on holistic raters. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: theoretical and empirical foundations* (pp. 237-265). Hampton Press.

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

Qualtrics. (2020). [Software]. https://www.qualtrics.com

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

Rea-Dickins, P., Kiely, R., & Yu, G. (2007). Student identity, learning and progression: The affective and academic impact of IELTS on 'successful' candidates. *IELTS Research Reports, 7*, 1-78.

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.

Riemenschneider, A., Weiss, Z., Schröter, P., & Meurers, D. (2021). Linguistic complexity in teachers' assessment of German essays in high stakes testing. *Assessing Writing, 50*. https://doi.org/10.1016/j.asw.2021.100561

Rifkin, B., & Roberts, F. D. (1995). Error gravity: a critical review of research design. *Language Learning, 45*(3), 511-537. https://doi.org/10.1111/j.1467-1770.1995.tb00450.x

Robitzsch, A., & Steinfeld, J. (2018). Item response models for human ratings: Overview, estimation methods, and implementation in R. *Psychological Test and Assessment Modeling, 60*(1), 101-139.

Roche, T., & Harrington, M. (2013). Recognition vocabulary knowledge as a predictor of academic performance in an English as a foreign language setting. *Language Testing in Asia, 3*(1), 1-13. https://doi.org/10.1186/2229-0443-3-12

Roche, T., & Harrington, M. (2014). Vocabulary knowledge and its relationship with EAP proficiency and academic achievement in an English-medium university in Oman. In R. Al-Mahrooqi & A. Roscoe (Eds.), *Focusing on EFL reading: theory and practice* (pp. 27-41). Cambridge Scholars Publishing.

Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual review of cognitive linguistics, 7*(1), 140-162. https://doi.org/10.1075/arcl.7.06rom

Römer, U. (2017). Language assessment and the inseparability of lexis and grammar: Focus on the construct of speaking. *Language testing, 34*(4), 477-492. https://doi.org/10.1177/0265532217711431

Rossiter, M. J., Abbott, M. L., & Kushnir, A. (2016). L2 Vocabulary research and instructional practices: Where are the gaps? *TESL-EJ, 20*(1), 1-25.

Ruegg, R., Fritz, E., & Holland, J. (2011). Rater sensitivity to qualities of lexis in writing. *TESOL Quarterly, 45*(1), 63-80. http://www.jstor.org/stable/41307616

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

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language testing, 25*(4), 465-493. https://doi.org/10.1177/0265532208094273

Schmitt, N. (1999). The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge. *Language testing, 16*(2), 189-216. https://doi.org/10.1177/026553229901600204

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Springer.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning, 64*(4), 913-951. https://doi.org/10.1111/lang.12077

Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (Ed.), *Formulaic Sequences: Acquisition, processing and use* (pp. 127-151). John Benjamins.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*(1), 26-43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition, 19*(1), 17-36. https://doi.org/10.1017/S0272263197001022

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, 47*(4), 484-503. https://doi.org/10.1017/S0261444812000018

Schmitt, N., & Zimmerman, C. B. (2002). Derivative word forms: What do learners know? *TESOL Quarterly, 36*(2), 145-171. https://doi.org/10.2307/3588328

Schoepp, K. (2018). Predictive validity of the IELTS in an English as a medium of instruction environment. *Higher Education Quarterly, 72*(4), 271-285. https://doi.org/10.1111/hequ.12163

Schoonen, R., van Gelderen, A., Stoel, R. D., Hulstijn, J., & de Glopper, K. (2011). Modeling the development of L1 and EFL writing proficiency of secondary school students. *Language Learning, 61*(1), 31-79. https://doi.org/10.1111/j.1467-9922.2010.00590.x

Schütze, C. T., & Sprouse, J. (2013). Judgment Data. In R. J. Podesva & D. Sharma (Eds.), *Research Methods in Linguistics* (pp. 27-50). Cambridge University Press.

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

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76*(1), 27-33. https://doi.org/10.1111/j.1540-4781.1992.tb02574.x

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

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Sinclair, J. M., Krishnamurthy, R., Daley, R., & Jones, S. (2004). *English collocation studies: the OSTI report*. Continuum.

Siyanova-Chanturia, A. (2015). On the 'holistic' nature of formulaic language. *Corpus linguistics and linguistic theory, 11*(2), 285-301. https://doi.org/doi:10.1515/cllt-2014-0016

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., & Omidian, T. (2020). Key issues in researching multiword items. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 529-544). Routledge. https://doi.org/10.4324/9780429291586-33

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., & Schmitt, N. (2008). L2 learner production and processing of collocation: a multi-study perspective. *The Canadian Modern Language Review, 64*(3), 429-458. https://doi.org/10.3138/cmlr.64.3.429

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). Multi-word 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

Skehan, P. (2009). Lexical performance by native and non-native speakers on language-learning tasks. In B. Richards, M. H. Daller, D. D. Malvern, P. Meara, J. Milton, & J. Treffers-Daller (Eds.), *Vocabulary studies in first and second language acquisition: the interface between theory and application* (pp. 107-124). Palgrave Macmillan UK. https://doi.org/10.1057/9780230242258_7

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.

Sonbul, S. (2015). Fatal mistake, awful mistake, or extreme mistake? Frequency effects on off-line/on-line collocational processing. *Bilingualism, 18*(3), 419-437. https://doi.org/10.1017/S1366728914000674

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of second language writing, 5*(2), 163-182. https://doi.org/10.1016/S1060-3743(96)90023-5

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *The language learning Journal, 36*(2), 139-152. https://doi.org/10.1080/09571730802389975

Thornbury, S. (2016). Communicative language teaching in theory and practice. In G. Hall (Ed.), *The Routledge Handbook of English language teaching* (pp. 224-237). Routledge.

Tidball, F., & Treffers-Daller, J. (2008). Analysing lexical richness in French learner language: what frequency lists and teacher judgements can tell us about basic and advanced words. *Journal of French Language Studies, 18*(3), 299-313. https://doi.org/10.1017/S0959269508003463

Tokowicz, N., & Kroll, J. (2007). Number of meanings and concreteness: Consequences of ambiguity within and across languages. *Language and cognitive processes, 22*. https://doi.org/10.1080/01690960601057068

Tomasello, M. (2000). The item-based nature of children's early syntactic development. *Trends in Cognitive Sciences, 4*(4), 156-163. https://doi.org/10.1016/S1364-6613(00)01462-5

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard 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

Uchihara, T., Eguchi, M., Clenton, J., Kyle, K., & Saito, K. (2021). To what extent is collocation knowledge associated with oral proficiency? A corpus-based approach to word association. *Language and Speech*. https://doi.org/10.1177/00238309211013865

Vafaee, P., & Suzuki, Y. (2020). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening ability. *Studies in Second Language Acquisition, 42*(2), 383-410. https://doi.org/10.1017/S0272263119000676

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

van Hout, R., & Vermeer, A. (2007). Comparing measures of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. 93-115). Cambridge University Press. https://doi.org/10.1017/CBO9780511667268.008

van Rijn, P. (2019). *Discussion of "Putting test scores on an interval scale"* [Powerpoint slides]. PIAAC seminar, Paris. https://www.oecd.org/skills/piaac/events/Item_9_%20Peter_van_Rijn.pdf

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-125). Ablex.

Vercellotti, M. L., Juffs, A., & Naismith, B. (2021). Multiword sequences in English language learners' speech: The relationship between trigrams and lexical variety across development. *System, 98*. https://doi.org/10.1016/j.system.2021.102494

Vermeer, A. (2004). The relation between lexical richness and vocabulary size in Dutch L1 and L2 children. In P. Bogaards & B. Laufer-Dvorkin (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 173-189). John Benjamins. https://doi.org/10.1075/lllt.10.13ver

Vilkaitė, L. (2016). Are nonadjacent collocations processed faster? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(10), 1632-1642. https://doi.org/10.1037/xlm0000259

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

Wallace, S. (2009). *A Dictionary of Education*. Oxford University Press. https://doi.org/10.1093/acref/9780199212064.001.0001

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.

Waring, R. (1997). A comparison of the receptive and productive vocabulary sizes of some second language learners. *Immaculata, 1*, 53-68.

Waring, R., & Browne, C. (n.d.). *The Online Graded Text Editor*. Extensive Reading Central. https://www.er-central.com/ogte/

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics, 28*(1), 46-65. https://doi.org/10.1093/applin/aml048

Webb, S. (2021). Word families and lemmas, not a real dilemma: Investigating lexical units. *Studies in Second Language Acquisition, 43*(5), 973-984. https://doi.org/10.1017/S0272263121000760

Webb, S., & Kagimoto, E. (2011). Learning collocations: Do the number of collocates, position of the node word, and synonymy affect learning? *Applied Linguistics, 32*(3), 259-276. https://doi.org/10.1093/applin/amq051

Webb, S., Newton, J., & Chang, A. (2013). Incidental learning of collocation. *Language Learning, 63*(1), 91-120. https://doi.org/10.1111/j.1467-9922.2012.00729.x

Weigle, S. C. (1994). *Effects of training on raters of English as a second language compositions: Quantitative and qualitative approaches*. ProQuest Dissertations Publishing.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language testing, 15*(2), 263-287. https://doi.org/10.1177/026553229801500205

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Wells, R. (1960). Nominal and verbal style. In T. A. Sebeok (Ed.), *Style in language* (pp. 213-220). MIT Press.

West, M. (1953). *A General Service List of English Words*. Longman, Green and Co.

Wilson, J., Roscoe, R., & Ahmed, Y. (2017). Automated formative writing assessment using a levels of language framework. *Assessing Writing, 34*, 16-36. https://doi.org/10.1016/j.asw.2017.08.002

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

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: measures of fluency, accuracy, & complexity*. Second Language Teaching & Curriculum Center, University of Hawai'i at Mānoa.

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

Wood, D. (2002). Formulaic language acquisition and production: Implications for teaching. *TESL Canada journal, 20*(1), 1-15. https://doi.org/10.18806/tesl.v20i1.935

Wood, D. (2020). Classifying and identifying formulaic language. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 30-45). Routledge. https://doi.org/10.4324/9780429291586-3

Wood, D., & Namba, K. (2013). Focused instruction of formulaic language: Use and awareness in a Japanese university class. *The Asian Conference on Language Learning Official Conference Proceedings 2013*, 203-212.

Wray, A. (2000). Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics, 21*(4), 463-489. https://doi.org/10.1093/applin/21.4.463

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

Wray, A. (2009). Conclusion: Navigating L2 collocation research. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 232-244). Springer. https://doi.org/10.1057/9780230245327

Wray, A. (2018). Concluding question: Why don't second language learners more proactively target formulaic sequences? In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding Formulaic Language: A Second Language Acquisition Perspective* (pp. 248-269). Routledge. https://doi.org/10.4324/9781315206615-13

Wray, A., & Namba, K. (2003). Use of formulaic language by a Japanese-English bilingual child: a practical approach to data analysis. *Japanese Journal for Multilingualism and Multiculturalism, 9*(1), 24-51.

Xerri, D., & Vella Briffa, P. (2018). Introduction. In D. Xerri & P. Vella Briffa (Eds.), *Teacher Involvement in High-Stakes Language Testing* (pp. 1-7). Springer International Publishing. https://doi.org/10.1007/978-3-319-77177-9_1

Xie, Q. (2019). Error analysis and diagnosis of ESL linguistic accuracy: Construct specification and empirical validation. *Assessing Writing, 41*, 47-62. https://doi.org/10.1016/j.asw.2019.05.002

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

Yoon, S.-Y., Bhat, S., & Zechner, K. (2012). Vocabulary profile as a measure of vocabulary sophistication. Proceedings of the 7th workshop on the innovative use of NLP for building educational applications, Montreal, Canada.

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics, 31*(2), 236-259. https://doi.org/10.1093/applin/amp024

Zhang, Y., & Elder, C. (2014). Investigating native and non-native English-speaking teacher raters' judgements of oral proficiency in the College English Test-Spoken English Test (CET-SET). *Assessment in education: principles, policy & practice, 21*(3), 306-325. https://doi.org/10.1080/0969594X.2013.845547

Zyzik, E. V. E., & Gass, S. (2008). Epilogue: A tale of two copulas. *Bilingualism, 11*(3), 383-385. https://doi.org/10.1017/S1366728908003611