Doctoral Thesis

Exploring the Role of Vocabulary in Writing: Predicting Lexical Diversity Scores, Differentiating Writing Proficiency, and Investigating Vocabulary Knowledge Development Over Time

Yajie Li

Division of Integrated Arts and Sciences
Graduate School of Integrated Arts and Sciences

Hiroshima University

September 2023

In loving memory of Guozhong Li

1937–2021

Summary

This dissertation investigates the role of vocabulary knowledge in L2 writing: the extent to which vocabulary knowledge scores can predict vocabulary use in writing activities for participants with different proficiency levels; how vocabulary knowledge scores can distinguish participant writing scores; and to what extent vocabulary knowledge scores and written production can track acquired vocabulary knowledge. To explore these questions, I conducted four experiments (three cross-sectional and one longitudinal) and distributed a range of vocabulary tasks to help determine participants' vocabulary knowledge.

Nation's framework of vocabulary knowledge (1990, 2001, 2013) divides vocabulary knowledge into receptive and productive dimensions. Receptive vocabulary knowledge requires participants to understand the form and meaning of words, and productive vocabulary knowledge demands participants to produce words. We can use vocabulary tasks as effective tools to test participants' vocabulary knowledge. Lexical diversity is one such effective tool to measure the variety of different words used in actual written texts or spoken production. By using lexical diversity measurements, it is possible to estimate vocabulary knowledge in writing use, writing competence, and overall language proficiency levels (e.g., Engber, 1995; Lu, 2012; Olinghouse & Leaird, 2009; Olinghouse & Wilson, 2013; Treffers-Daller, 2013; Treffers-Daller et al., 2018; Vidal & Jarvis, 2020; Yu, 2010).

The first experimental chapter partially replicates Treffers-Daller et al. (2018) and explores potential relationships between vocabulary tasks and L2 written production for participants at the Common European Framework of Reference (CEFR) A2 level (Council of Europe, 2001). It examines whether vocabulary scores can predict participants' vocabulary in writing use with 29 L1 Chinese participants. I gave participants four vocabulary knowledge tasks: Lex30, a task based on word association (Meara & Fitzpatrick, 2000); G_Lex, a single-word gap-fill task (Fitzpatrick & Clenton, 2017); the Productive Vocabulary Levels Test (the

PVLT), a sentence completion task (Laufer & Nation, 1999); the Vocabulary Levels Test (the VLT), a form-meaning matching task (Nation, 1983; Schmitt, 2000) assessing receptive vocabulary knowledge; and one writing topic (see Appendix A for specific examples of Lex30, G_Lex, the PVLT, the VLT, and writing topic).

The second experimental chapter focuses on productive vocabulary knowledge tasks and investigates potential relationships between productive vocabulary tasks and L2 written production for participants at CEFR levels B1 to C1. This experiment examines 91 L1 Japanese participants with higher proficiency levels. Considering that writing is a productive skill and that the PVLT task also accesses facets of receptive vocabulary knowledge (Edmonds et al., 2022; Webb, 2008), I gave participants three productive vocabulary tasks (Lex30, G_Lex, the PVLT) and one IELTS writing topic (see Appendix B for sample responses of Lex30, G_Lex, the PVLT, and IELTS writing).

The third experimental chapter examines how productive vocabulary tasks can differentiate between IELTS writing scores. 63 L1 Japanese speakers and 35 L1 French speakers participated in this experiment. All participants finished the three productive vocabulary tasks (Lex30, G_Lex, and the PVLT) and two IELTS writing topics (see Appendix E for sample responses of Lex30, G_Lex, the PVLT, and IELTS writing). Qualified IELTS raters marked all the writing samples based on the IELTS writing rubric (see Appendix C for IELTS writing band descriptors). I divided all participants from the two different language backgrounds into different proficiency groups based on their IELTS writing scores.

The fourth experimental chapter explores how productive vocabulary knowledge task scores and lexical diversity measure scores relate over a short study period. It investigates whether participants' vocabulary knowledge and lexical diversity scores can improve through a pre- and post-test design over a short-term intervention (approximately 12 weeks).

Participants from a single language background (L1 Japanese participants, N=51) with similar proficiency levels joined the current experiment. I used two versions of three productive vocabulary knowledge tasks (Lex30, G_Lex, and the PVLT) and two IELTS writing topics (with different question prompts at each test time) at the beginning and the end of the study period (see Appendix F for sample responses of vocabulary tasks and IELTS writing at testing time one and testing time two). I gave all participants the same vocabulary lists (2K New General Service List [NGSL]) to learn.

In a partial replication of Treffers-Daller et al.'s (2018) study, I use the lemma, which comprises 'a headword and its inflected forms' (Nation, 2016, 2022), as word unit both for responses from vocabulary tasks and writing samples in my first experiment. A research gap in Treffers-Daller et al.'s study left using the flemma, which comprises 'a headword and inflected forms of different parts of speech' (Nation, 2016, 2022) as a word unit unexplored, I have used the flemma as a word counting unit for my second, third, and fourth experiments.

The findings from the four experimental chapters raise several key issues to discuss and explore further: (i) disparities in accessing vocabulary knowledge used in written production, as evidenced by vocabulary knowledge measures: vocabulary knowledge measures differ in task features and task embeddedness; (ii) measuring vocabulary knowledge can differentiate levels of proficiency in IELTS writing; (iii) selecting appropriate measures of lexical diversity depends on the specific research question or goal, as different measures may have different strengths and limitations. Traditional lexical diversity measure scores show greater accuracy with vocabulary knowledge scores in writing use, whereas the more recently devised lexical diversity measures show better performance in tracking vocabulary knowledge development in written production; (iv) G_Lex shows greater power in tracking vocabulary knowledge improvement than the PVLT and Lex30; and, (v) using

online flashcards learning with 2K NGSL lemma-based word lists offers an effective means to improve vocabulary knowledge and vocabulary in writing use.

In conclusion, I hope the issues identified and investigated regarding the dynamic relationship between vocabulary knowledge and written production can support future research. The findings' implications are significant for L2 writing class and vocabulary knowledge assessment.

Acknowledgements

First, I would like to thank my primary academic advisor, Dr. Jon Clenton, who provided invaluable guidance, patience, and support throughout my PhD studies. He has opened the door to conducting vocabulary research and inspired and encouraged me to become independent.

I sincerely thank Professor Simon Fraser for his guidance, expertise, and support for my PhD studies. I also thank Dr. George Higginbotham for his help in my research.

I sincerely thank my dissertation committee, Dr. Noriko Yamane and Dr. Masahiro Shinya, for their time and insightful comments on my dissertation. I would like to thank Professor Katsumi Iwasaki for his suggestions for this dissertation. I am grateful to Professor Itaru Nagasaka and Professor Carolin Funck for their input on my dissertation.

I extend my heartfelt gratitude to Professor Jeanine Treffers-Daller for her thoughtful suggestions in support of my studies. Special thanks also go to Dr. Yan Zhao for his unwavering support throughout the process. I thank Professor Dylan Jones, who proofread the whole dissertation.

I acknowledge the help from Dr. TJ Boutorwick, Dr. Hosam Elmetaher, Dion Clingwall, and Yu Wang. I thank the participants for their kind engagement in my studies.

Finally, I express profound gratitude to my family for providing me with the strength to complete this PhD study.

Table of Contents

List of Tables

List of Figures

Glossary

Term	Explanation
Word (counting) unit	The lexical unit comprising words. The most common terms of word units include tokens (the total number of words in a text), word types (the number of unreproduced words), word families, lemmas, flemmas, or other levels of word families preferred by researchers.
Lemma	Lemma means a headword with its inflected forms of the same part of speech. If we use lemma as a word unit, the adjective *abstract*, noun *abstract/abstracts*, and verb *abstract/abstracts/abstracted/abstracting* would be counted as three different words (lemmas). Lemma count assumes that learners have the word knowledge of inflected forms but do not have the part-of-speech knowledge.
Flemma	Flemma is similar to lemma, but do not distinguish part-of-speech of words. If we use flemma as a word unit, the word (*abstract*) comprises adjective (*abstract*), noun (*abstract/abstracts*), and verb (*abstract/abstracts/abstracted/abstracting*) would be counted as one word (flemma). Flemma count assumes that learners have the word knowledge of inflected forms and can distinguish the part of speech of words.
Word Family	Seven different levels were proposed by Bauer and Nation (1993). Word families consist of a headword with its inflected forms and the most derived forms. If we use word family as a word unit, the inflected forms of abstract (*abstract*, *abstracts*, *abstracting*, *abstracted*) and derived forms of abstract (*abstractedly*, *abstractly*, *abstractness*, *abstraction*, *abstractions*) would all be counted as the same word. Word family count assumes that learners have the knowledge of inflected forms and derived forms of the words.
Word Family Level 1	A different form is a different word. Capitalization is ignored.
Word Family Level 2	Regularly inflected words are part of the same family. The inflectional categories are - plural; third person singular present tense; past tense; past participle; *-ing*; comparative; superlative; possessive.
Word Family Level 3	*-able* (makes adjectives from verbs, adding the meaning "able to be ~ed" where the swung dash is the verb in the stem: *acceptable*), *-er* (makes nouns from verbs: *computer*), *-ish* (adjectives from nouns, numbers and adjectives: *selfish*), *-less* (makes adjectives from nouns adding the meaning 'less, without': *useless*), *-like* (makes adjectives from nouns, meaning 'resembling ~': *businesslike*), *-ly* (makes adverbs from adjectives: *probably*), *-ness* (makes nouns from adjectives: *goodness*), *-th* (makes ordinal numbers: *sixth*), *-y* (makes adjectives from nouns: *funny*), *non-* (makes negatives with nouns and adjectives: *nonstop*), *un-* (makes negatives with adjectives and adverbs: *unclear*), all with restricted uses.
Word Family Level 4	*-al* (makes adjectives from nouns: *national*), *-ation* (makes nouns from verbs: *information*), *-ess* (female nouns: *princess*), *-ful* (makes adjectives from nouns adding the meaning 'full': *beautiful*), *-ism* (makes nouns describing a way of thinking or belief: *nationalism*), *-ist* (makes nouns describing a person with a particular

	belief or job: *artist*), *-ity* (makes nouns from adjectives: *security*), *-ize* (makes verbs: *realize*), *-ment* (makes nouns from verbs: *government*), *-ous* (makes adjectives: *dangerous*), *in-* (negative: *inability*), all with restricted uses.
Word Family Level 5	*-age* (makes nouns from verbs: *leakage*), *-al* (makes nouns from verbs: *arrival*), *-ally* (makes adverbs: *idiotically*), *-an* (makes nouns showing a job or regional origin: *American*), *-ance* (makes nouns from verbs: *clearance*), *-ant* (makes nouns from verbs: *consultant*), *-ary* (makes adjectives: *revolutionary*), *-atory* (makes adjectives from verbs: *confirmatory*), *-dom* (makes nouns: *kingdom; officialdom*), *-eer* (person: *black marketeer*), *-en* (makes adjectives from nouns: *wooden*), *-en* (makes verbs from adjectives: *widen*), *-ence* (makes nouns from verbs: *emergence*), *-ent* (makes adjectives from verbs: *absorbent*), *-ery* (nouns usually indicating a collection or group: *bakery; trickery*), *-ese* (makes nouns indicating an inhabitant or language: *Japanese; officialese*), *-esque* (added to proper names indicating a style: *picturesque*), *-ette* (marking small size: *usherette; roomette*), *-hood* (indicating a state of being: *childhood*), *-i* (indicating nationality: *Israeli*), *-ian* (largely added to proper nouns indicating inhabitants, places of regional origin, and languages or to common nouns to indicate jobs: *phonetician; Johnsonian*), *-ite* (added to proper nouns to indicate an inhabitant or supporter: *Trotskyite*; also chemical meaning), *-let* (nouns meaning "little": *coverlet*), *-ling* (nouns largely indicating a young animal: *duckling*), *-ly* (makes adjectives: *brotherly*), *-most* (adjectives indicating extreme: *topmost*), *-ory* (makes adjectives from verbs: *contradictory*), *-ship* (nouns indicating a state of being: *studentship*), *-ward* (makes adverbs indicating direction: *homeward*), *-ways* (makes adverbs indicating direction: *crossways*), *-wise* (makes adverbs indicating manner or 'from the point of view of': *endwise; discussion-wise*), *anti-* (against: *anti-inflation*), *ante-* (before: *anteroom*), *arch-* (most important: *archbishop*), *bi-* (two: *biplane*), *circum-* (around: *circumnavigate*), *counter-* (in opposition to: *counter-attack*), *en-* (verbs from nouns: *encage; enslave*), *ex-* (out, moving away: *ex-president*), *fore-* (in front of: *forename*), *hyper-* (too much or very large: *hyperactive*), *inter-* (between, back and forth: *inter-African, interweave*), *mid-* (middle: *mid-week*), *mis-* (wrong: *misfit*), *neo-* (new: *neo-colonialism*), *post-* (after: *post-date*), *pro-* (in favour of: *pro-British*), *semi-* (half: *semi-automatic*), *sub-* (under: *subclassify; subterranean*), *un-* (with verbs indicating reversal of an action: *untie; unburden*).
Word Family Level 6	*-ible* (makes adjectives, a version of -able: *forcible*), *-ee* (a person who is ~ed: *employee*), *-ic* (makes adjectives: *basic*), *-ify* (makes verbs: *simplify*), *-ion* (makes nouns: *education*), *-ition* (makes nouns: *addition*), *-ive* (makes adjectives: *expensive*), *-th* (makes nouns: *truth*), *-y* (makes nouns: *safety*), *pre-* (before: *preschool*), *re-* (again: *reunify*).
Word Family Level 7	Classical roots and affixes

Word Family (WF) 6	The term WF6 is used in McLean's (2018) article, which excludes level 7 based on the criteria of Bauer and Nation (1993).
Lexical diversity measure(s)	Lexical diversity evaluates the distribution range of words variety in writing or speaking contexts. The lexical diversity measures used in the current dissertation include 11 lexical diversity measures including word types, TTR, Root_TTR, Log_TTR, MSTTR, MAAS, D(vocd), HD-D, MTLD, MTLD_W, and MATTR, which illustrates below.
Word type	The occurrence of unique words in a text would be counted as different words.
Type-token Ratio (TTR)	The number of different word types divided by the number of tokens (the total number of words in a text) (Johnson, 1944).
Root_TTR	Also known as Guiraud's index (see Guiraud, 1954). Root_TTR shows the ratio between types and the square root of tokens. $\text{Root\_TTR}=\frac{\text{Types}}{\sqrt{\text{Tokens}}}$
Log_TTR	Also referred to as Herdan's index *C*. Log_TTR (Herdan, 1960) means the number of log types divided by the log tokens. $\text{Log\_TTR}=\frac{\text{logTypes}}{\text{logTokens}}$
Mean segment type-token ratio (MSTTR)	MSTTR (Johnson, 1944) divides the text into several segments and calculating the average TTR scores for the segments. The current dissertation uses 50 words as a segment.
MAAS	MAAS index is based on the logarithmic curve (Maas, 1972). Maas index: $a^2=\frac{\text{LogTokens-LogTypes}}{\text{Log}^2\text{Tokens}}$
D(vocd)	D (vocd) measure estimates a random sampling process of texts, selecting 35 tokens from a random sample of 100 words and then moving from 36 tokens to 50 tokens (Malvern & Richards, 1997). $\text{TTR}=\frac{D}{N}\left[\left(1+2\frac{N}{D}\right)^{\frac{1}{2}}-1\right]$
The hypergeometric distribution of D (HD-D)	HD-D index chooses a 42-word random sample and then computes the chances that every token can be met in this random sample. It is an index based on the hypergeometric distribution (McCarthy & Jarvis, 2007, 2010)
The measure of textual lexical diversity (MTLD)	MTLD is a measure of *textual* lexical diversity (McCarthy, 2005). MTLD insists on a fixed TTR value (e.g., 0.72) and computes the TTR from the first word, the first two words, and the adding one word at a time until the TTR falls below 0.72.
MTLD Wrap around (MTLD_W)	MTLD-W (Vidal & Jarvis, 2020) uses the moving window method (the same as MATTR, explained below) and a wrap-around process to compute the final segment by forwarding the last part of a text by adding words from the end to the beginning of the text until it reaches a 0.72 value.

Moving average type-token ratio (MATTR)	MATTR (Covington & McFall, 2010) uses a moving window method, such as taking 50 tokens as a segment of a text from the beginning until it reaches the last token of the text. The final MATTR value is the mean value of all segments.

Note. Summary of the levels of word families and its explanations and example words adapted from L. Bauer and P. Nation (1993) "Word Families", *International journal of Lexicography*, *6*(4), as cited in "What is morphological awareness and how can you develop it?" by P. Nation and L. Bauer, 2023, *Language Teaching Research*, *33*, p. 83. For a more detailed explanation about units of word counting including types, lemmas, flemmas, word families, see Nation (2016) "Making and Using Word Lists for Language Learning and Testing." *John Benjamins Publishing Company*, p. 12–13, p. 23–27.

Chapter 1: Introduction

1.1 Introduction

The essential role of vocabulary acquisition for second language (L2) learners of English is widely acknowledged (e.g., Grabe & Stoller, 1997; Jiang, 2002; Nation, 2001, 2013; Read, 2000; Schmitt, 2008). L2 English language learners with a larger vocabulary knowledge size have higher language proficiency. We can evaluate language learner proficiency levels in the four language skills: listening, reading, speaking, and writing. Previous studies have shown the relationships between vocabulary knowledge and listening (e.g., Chang, 2007; Zhang & Graham, 2020); vocabulary knowledge and reading (e.g., Masrai, 2019; Zhang & Annual, 2004); and vocabulary knowledge and speaking (e.g., Clenton et al., 2020; de Jong et al., 2012). Writing is often used in large-scale testing suites, and a recent study by Treffers-Daller et al. (2018) showed a significant correlation between vocabulary scores and writing proficiency.

This dissertation follows up on this important thread: examining the relationship between vocabulary and writing can provide important guidance for language assessment and language pedagogy. To provide full details regarding the construct needed, it is imperative to assess both vocabulary knowledge and writing proficiency. Thus, the importance of vocabulary knowledge in one of these skills, writing, is the focus of the current dissertation. This dissertation explores the important role of vocabulary knowledge in L2 writing proficiency levels and to what extent deliberate vocabulary learning can improve L2 language learners' vocabulary knowledge in use. The current dissertation uses a range of assessment tools. Specifically, one useful tool for assessing L2 language learners' vocabulary knowledge is vocabulary tests, which provide a quick and useful way to evaluate vocabulary knowledge. Vocabulary tests, designed based on different aspects of vocabulary knowledge, provide immediate feedback on L2 learners' performance.

To gain a deeper understanding of vocabulary, it is important to remember that vocabulary grows over time (e.g., Huang, 2010; Zhong & Hirsh, 2009). A vital component of language learning is to assess learners' current vocabulary knowledge as well as to provide them with opportunities to expand their vocabulary going forward. We can expand L2 learners' vocabulary knowledge both deliberately and incidentally. According to Pellicer-Sánchez (2020), '(t)he use of vocabulary activities that explicitly direct learners' attention to unknown lexical items creates the conditions for deliberate learning to occur' (p. 159), whereas 'meaning-focused activities with which learners engage for communicative purposes, without a specific intention to learn new vocabulary, create the conditions for incidental learning to occur' (p. 183). As Nation (2020) showed, 'word knowledge develops over a period of time' and 'vocabulary knowledge is most likely to develop if there is a balance of incidental and deliberate appropriate opportunities for learning' (p. 15).

One of the objectives of the current dissertation, therefore, is to examine how deliberate vocabulary learning, using word lists through flashcard learning activities, can improve vocabulary knowledge. To explore this aim, I employ vocabulary knowledge tests to track the vocabulary knowledge of L2 participants.

To explore these issues, I conduct three cross-sectional studies and one longitudinal study in the present dissertation. In my cross-sectional studies, I explore the relationships between vocabulary knowledge task scores and vocabulary in written context use by analysing lexical diversity scores. I include participants of different proficiency levels. I also examine the extent to which vocabulary task scores can differentiate between writing levels. In my longitudinal study, I investigate whether participants' vocabulary knowledge and vocabulary knowledge in written contexts can improve through a pre- and post-test design. Participants learn words from a word list each week. They can learn the words by using

digital flashcards on their electronic devices or by participating in in-class warm-up activities provided by their instructors.

1.2 Background

With an increasing number of students taking high-stakes exams and entering universities, providing instant feedback on their vocabulary knowledge is proving increasingly pertinent. Because high-stakes language exams, such as the International English Language Testing System (IELTS), require test-takers to have a certain level of vocabulary, with vocabulary being one of the scoring criteria, the central place for vocabulary knowledge in testing domains and student knowledge is clear. Vocabulary knowledge is an important component for academic success required for different language skills (as shown above), and language acquisition (e.g., Alqahtani, 2015; Nagy & Townsend, 2012; Webb & Chang, 2012).

Indicators, such as the Common European Framework of Reference (CEFR) (Council of Europe, 2001), suggest that vocabulary is central to determining proficiency. Research by Treffers-Daller et al. (2018) has supported this view of the central place for vocabulary in CEFR placement. Treffers-Daller et al. (2018) outlined the extent to which various vocabulary knowledge measures in context appear to predict student proficiency levels, ranging from CEFR B1 to C2. Their investigation was based on relationships between lexical diversity (LD) scores and overall CEFR levels. However, they used the CEFR as a composite measure, so, since it includes all four skills (i.e., reading, listening, speaking, and writing), the possibility remains that Treffers-Daller et al.'s data were not indicative only of the learners' writing skills. I detail this point in chapter 3, section 3.2.

To avoid the issues related to Treffers-Daller et al.'s study, this dissertation focuses on the relationship between vocabulary knowledge and writing. A further distinction from

Treffers-Daller et al.'s paper is their use of vocabulary measures. Their study reported significant correlations between vocabulary scores and writing levels. Their vocabulary scores were provided by high-stakes proficiency test agencies based on a range of discrete variables. This motivated the current dissertation to investigate how multiple vocabulary knowledge measure scores relate to writing levels.

I hope the low-stakes vocabulary knowledge measures validated in the current dissertation may clarify some vocabulary knowledge/language assessment issues. Milton (2009) highlighted how 'low-stakes testing … might provide the same information at far less cost, effort and disruption to the education process for schools and learners' compared to 'the full panoply of the formal examination system', and 'vocabulary sizes can help suggest much more appropriate CEFR levels' (pp. 191–192). Taking high-stakes English language tests, such as the International English Language Testing System (IELTS) test, is costly for many language learners who thus may struggle to know their language proficiency but want to improve their English language proficiency levels at different stages. Meanwhile, the high-stakes English language test may stop being widely offered by *force majeure* under unforeseen circumstances. Over the past few years, during the pandemic period, high-stakes language tests may have been interrupted around the world, causing inconvenience to test-takers who want to assess their language proficiency. The concern also exists that 'the switch to accepting at-home proficiency tests for high-stakes decisions raises many concerns for stakeholders, such as technological demands, exam security, and validity of score use' (Isbell & Kremmel, 2020, p. 600).

1.3 Vocabulary Knowledge in L2 Written Production

Studies have shown that vocabulary is important to language proficiency (Daller & Phelan, 2013; Roche & Harrington, 2013; Qian & Lin, 2019; Stæhr, 2008; Trenkic &

Warmington, 2019; Zareva et al., 2005; Zhang & Zhang, 2022). For example, Zhang and Zhang (2022) have reported that 'true' correlations between vocabulary knowledge and reading or listening fall within the range of .56 and .67 and that vocabulary knowledge accounts for 31%–45% of the variance in L2 comprehension. Relationships between vocabulary and use depend on a broad array of linguistic resources, and a clear understanding and command of requisite vocabulary allows users to express themselves accurately and concisely (Schoonen et al., 2011). Importantly, research has indicated that language learners need to develop the vocabulary to be successful in high-stakes assignments (Coxhead, 2012), and 'vocabulary is consistently identified as the best predictor of academic success for EFL (English as a foreign language) students in HE (higher education)' (Trenkic & Warmington, 2019, p. 363).

Studies have shown that vocabulary knowledge positively correlates with writing competence and that learners with greater vocabulary knowledge perform better or acquire higher competency in written production than counterparts with lower vocabulary knowledge (Henriksen & Danelund, 2015; Johnson et al., 2016; Kiliç, 2019; Laufer & Nation, 1995; Milton et al., 2010; Roche & Harrington, 2013; Treffers-Daller et al., 2018). We can examine vocabulary knowledge using a variety of vocabulary tests, and vocabulary test results can predict learners' achievement in their written production (e.g., Henriksen & Danelund, 2015; Johnson et al., 2016; Kiliç, 2019; Laufer & Nation, 1995; Milton et al., 2010; Roche & Harrington, 2013). Some previous studies examining the relationship between vocabulary knowledge tests and writing proficiency have usually used a single receptive vocabulary test (e.g., Milton et al., 2010; Roche & Harrington, 2013; Treffers-Daller et al., 2018) or a single productive vocabulary test (e.g., Laufer & Nation, 1995). In addition, other studies have used multiple vocabulary knowledge tests to explore the relations between vocabulary knowledge

and writing performance, mainly through human raters' judgement on writing performance and frequency-based computation process (e.g., Johnson et al., 2016; Kiliç, 2019).

To my knowledge, though, only two studies (Henriksen & Danelund, 2015; Laufer & Nation, 1995) have investigated the relations between vocabulary knowledge and writing using vocabulary knowledge tests and lexical richness measures. Laufer and Nation (1995) used the Productive Vocabulary Levels Test (PVLT) and the Lexical Frequency Profile (LFP) to explore the relations between vocabulary knowledge and vocabulary knowledge in written use. Henriksen and Danelund (2015) examined the vocabulary knowledge of secondary school participants through three vocabulary tests: the Vocabulary Levels Test (VLT, Nation, 1983); the PVLT (Laufer & Nation, 1995, 1999); and Lex30 (Meara & Fitzpatrick, 2000), as well as the four lexical richness measures of tokens (the number of all running words in a language sample), type/token ratios (measuring the number of unique words/types), Guiraud index (the number of types divided by a square root of tokens), and Advanced Guiraud (calculating types above 2K, and using the same formula as Guiraud index). Laufer and Nation's study claimed to use lexical richness measures, but in fact, they employed a frequency-based method (the LFP). Thus, Henriksen and Danelund's study is the only study so far to use both vocabulary tests and lexical diversity measures to explore the relations between vocabulary knowledge and writing. The findings in Laufer and Nation's study and Henriksen and Danelund's study showed that participants with larger vocabulary knowledge also present greater lexical variation in their written production.

1.3.1 The Importance of Using Vocabulary Tests and Lexical Diversity Measures to Assess Vocabulary Knowledge and L2 Written Production.

The construct of what is a 'word' is difficult to define (e.g., Nation 2013; Gardner, 2007), and vocabulary knowledge dimensions involve both quality ('depth') and vocabulary

size (‘breadth’) (Anderson & Freebody, 1981; Read, 2000). Some scholars (Meara, 1990; Corson, 1995; Laufer, 1998) have used *passiv*e and *active* for receptive and productive. Read (2000, pp. 155–156) employed the terms *recognition* and *comprehension* for receptive and *recall* and *use* for productive. Meanwhile, Nation (1990, 2001, 2013) indicated that receptive–productive is a major vocabulary knowledge scale. He made distinctions among *form*, *meaning* and *use* by combining receptive and productive sides to explain different dimensions of word knowledge. The distinction between receptive and productive usually refers to the receptive skills of listening and reading and the productive skills of speaking and writing (e.g., Palmer, 1921). Among all these different definitions of vocabulary knowledge, Nation’s dimensions of vocabulary knowledge represent one of the most well-known frameworks in the research community.

Most research to date has focused on receptive vocabulary knowledge rather than productive vocabulary measures. Receptive vocabulary tests require participants to recognize the form or the meaning of the words. In contrast, productive vocabulary knowledge tests measure vocabulary in use by requiring test-takers to produce vocabulary. A potential reason for this balance and bias in the research might be that testing productive vocabulary knowledge is reported as being more difficult than accessing and assessing receptive vocabulary knowledge (e.g., Nation, 2013; Schmitt, 2014, 2019). Previous studies have highlighted the multidimensional feature of testing vocabulary knowledge (e.g., Chapelle, 2006; Laufer, 1998; Nation, 2007). Because of the multidimensional feature and there being no tests that can tap all vocabulary knowledge dimensions, empirical studies usually examine vocabulary knowledge through multiple measures (e.g., Fitzpatrick, 2007; Fitzpatrick & Clenton, 2017). Following previous studies, the current dissertation also uses multiple vocabulary knowledge measures to assess vocabulary knowledge.

In addition, studies have used lexical diversity measures to distinguish between proficiency levels and predict language learners' general language ability (e.g., Treffers-Daller et al., 2018; Treffers-Daller, 2013; Vidal & Jarvis, 2020; Jarvis & Hashimoto, 2021; Lu, 2012). Studies have also treated lexical diversity measures as predictors to forecast writing proficiency levels (e.g., Engber, 1995; Jarvis, 2002; Olinghouse & Leaird, 2009; Olinghouse & Wilson, 2013; Yu, 2010). However, upon further examination, it becomes apparent that many constructs are multidimensional, with the lexical diversity construct still under development and needing further refinement (Jarvis, 2013a; Kim et al., 2018). Helpfully, Jarvis (2013a, 2013b) has proposed six features of LD measures: (i) variability (inherent property of redundancy), (ii) volume (vocabulary size), (iii) evenness (balance), (iv) rarity (less common/frequent words), (v) dispersion (spatial distribution), and (vi) disparity (degree of differentiation). The current dissertation thus follows Jarvis's definition of lexical diversity and includes a variety of LD measures.

1.4 Measuring Vocabulary Knowledge Development Through Deliberate Vocabulary Learning

Measuring vocabulary knowledge development is important for L2 learners and language instructors to encourage effective learning and to strive for pedagogical improvement (Nation, 2020; Schmitt, 2019). Conducting longitudinal studies can reveal how vocabulary knowledge can be gained through various intervention measures. It can help language learners identify their vocabulary knowledge gaps and focus on the vocabulary knowledge they need to improve. Language instructors can also assess the effectiveness of teaching methods and curriculum and make adjustments to improve their classroom teaching materials and practice.

However, there is a lack of longitudinal studies in the vocabulary knowledge research community, as identified by Pellicer-Sánchez (2019). Earlier studies investigated vocabulary development based on a single test (Cobb & Horst, 2001; Fitzpatrick, 2012). Daller et al. (2013) investigated writing level development through lexical diversity measures and human ratings. One of the current dissertation's aims is to explore the extent to which vocabulary knowledge can be acquired over a short study period. To investigate how vocabulary knowledge develops over time, I focus on improving participants' vocabulary knowledge. The current dissertation examines to what extent participants' vocabulary knowledge and their vocabulary knowledge in writing use can be developed using deliberate word list learning activities. I use multiple vocabulary knowledge and lexical diversity measures to track changes in vocabulary knowledge and writing proficiency.

1.5 Overview of the Dissertation

This study aims to explore the relationship between vocabulary knowledge and L2 written production and identify how vocabulary knowledge tasks can track development. I use multiple vocabulary tasks to assess participant vocabulary knowledge and a range of lexical diversity measures to evaluate vocabulary used in written production.

Chapter 2 presents the literature review containing three main sections. The first section presents a review of vocabulary knowledge task studies; the second section examines a review of lexical diversity measure studies; and the third section provides a review of word counting unit studies. Chapters 3–5 report on three empirical studies that explore the relations between vocabulary knowledge and IELTS written production. These three experimental chapters explore how vocabulary knowledge scores can predict lexical diversity scores in writing for participants of different proficiency levels and how vocabulary knowledge tasks can distinguish between different IELTS writing levels as judged by qualified raters. Chapter

6 evaluates to what extent vocabulary knowledge tests can track vocabulary knowledge developed over time. Chapter 7 discusses the various threads based on the findings from the experimental chapters, tying them together and synthesising them in terms of the implications for both pedagogical practice and future research.

Chapter 2: Literature Review

2.1 Introduction

This dissertation investigates how vocabulary knowledge influences the written work of L2 English language learners. To explore this topic, I examine how vocabulary knowledge can be assessed and the extent to which vocabulary knowledge tasks can reflect participant vocabulary knowledge through validated vocabulary tasks. In addition, I employ lexical diversity measures which have long been used to predict several facets of written vocabulary production. Specifically, lexical diversity measures are used to predict vocabulary size, vocabulary knowledge proficiency, writing proficiency, human judgement writing scores, and general language proficiency levels. In light of this, the current dissertation employs lexical diversity measures to evaluate participants' vocabulary knowledge used for their writing. Thus, the current literature review chapter provides literature reviews from three standpoints: (i) vocabulary knowledge task studies, (ii) lexical diversity measure studies, and (iii) word counting unit studies. The literature review also provides a foundation for my experimental chapters (chapters 3–6).

The following literature review chapter comprises five sections. The first section (section 2.2) reviews five vocabulary knowledge task studies. All these vocabulary knowledge studies focus on productive vocabulary knowledge tasks. One study, from Laufer and Nation (1995), explored the relations between one productive vocabulary knowledge task and lexical richness in participants' written production. The papers reviewed in the current chapter include several published productive vocabulary knowledge tasks that will be used for my experimental chapters.

The second section (section 2.3) summarises and synthesises papers relating to lexical diversity measures. I select four papers that validate lexical diversity scores focusing on exploring lexical diversity measures with vocabulary scores, writing levels, and general

language proficiency levels. In addition, the four papers reviewed in the second section raise one of the main issues concerning the fact that different word counting units can cause different lexical diversity scores.

I then present my third literature review section (section 2.4), which explores the particular problems with using various word counting units. This section analyses two papers that have explored the appropriate word counting units for L2 English language learners and the suitable word counting units for lexical diversity measures.

The fourth and fifth sections (section 2.5 and section 2.6) summarise the pertinent points within papers in the literature associated with productive vocabulary knowledge tasks, lexical diversity measures, and word counting units, and thus extrapolates and presents an outline of research questions for the experimental chapters that follow.

2.2 Review of Vocabulary Knowledge Task Studies

Five papers selected in this review section discuss vocabulary knowledge tasks, which represent landmark studies in what they proposed or validated. The section starts with Laufer and Nation's (1995) paper. Their paper investigated how lexical richness was manifest in participants' written production using the Lexical Frequency Profile (LFP). They assessed participants' vocabulary knowledge through an active version of the Vocabulary Levels Test (Nation, 1983), which was later renamed the Productive Vocabulary Levels Test (PVLT) in Laufer and Nation's later (1999) paper. Laufer and Nation's (1999) paper suggested that the PVLT can predict participants with different proficiency levels. Citing issues with such testing, Meara and Fitzpatrick (2000) presented the Lex30 task as a more effective method when assessing productive vocabulary knowledge with fewer contextual demands than the PVLT. Walters (2012) further validated the Lex30 task with the PVLT, and a translation test, but also proposed the recall/use issues related to Lex30. Fitzpatrick and Clenton (2017) validated four productive vocabulary knowledge tasks (Lex30, G_Lex, the LFP, and the BFP), and proposed a vocabulary knowledge capture model encompassing the four tasks used in their study.

2.2.1 Laufer, B., & Nation, P. (1995): Vocabulary Size and Use: Lexical Richness in L2 Written Production.

Laufer and Nation's (1995) study proposed an innovative means to examine lexical richness in students' writing termed the Lexical Frequency Profile (LFP), which broke down learners' essays in terms of lexical frequency using a computer program. Their results indicated that the LFP might be a trustworthy measure in examining lexical richness by exploring the stability in writing for two topics with an identical set of L2 learners. They also claimed that the LFP measure could differentiate between proficiency levels and reflect

learners' vocabulary quantity in use citing positive correlations between the LFP and the active version of the Vocabulary Levels Test (Nation, 1983). They emphasised that the LFP was a helpful measure in evaluating both writing quality and vocabulary development.

Laufer and Nation mentioned that vocabulary size was crucial in determining writing quality, especially for L2 learners with a limited vocabulary size when compared to L1 English speakers. Lexical richness measurements were used in accessing vocabulary quality relating to variety and size and distinguishing the relationships between vocabulary knowledge and vocabulary size. In Laufer and Nation's paper, they cited Engber (1993), who reported a positive correlation of .57 between lexical variation and writing quality. Engber's work indicated that vocabulary knowledge in active use helped with writing quality. Laufer and Nation outlined the relationships between vocabulary size and use, imperative for exploring learners' vocabulary knowledge use when they are required to produce lexis. Laufer (1991) evaluated lexical richness knowledge with lower-level learners over fourteen and twenty-eight weeks. In Laufer's (1991) study, it was unclear whether the vocabulary development resulted from learning new vocabulary or was related to activating previously learned vocabulary. To show learners' vocabulary size, Laufer and Nation (1995) proposed a new measure, the aforementioned LFP, which reflects vocabulary size. They assumed that learners with an extensive vocabulary could produce a higher quality of writing that showcased their vocabulary knowledge.

Laufer and Nation (1995) introduced and identified several measures to describe lexical richness, including lexical originality (LO), lexical density (LD), lexical sophistication (LS), lexical variation (LV), semantic variation (Mendelson, 1981), lexical quality (Arnaud, 1984; 1992), T-unit length, and error-free T-unit length (Cohen, 1989). Lexical originality described the percentage of unique words written by one learner writing, which could easily affect different writers or topics. Lexical density was the ratio between the number of lexical

words (e.g., nouns, verbs, adjectives, and adverbs) and total tokens. It would be influenced by the number of function words used to reflect the structures of writings. Lexical sophistication related to the ratio between advanced tokens and tokens, but different researchers determined the different advanced lexis standards, which were unstable and lack of a consensus in deciding advanced tokens in practical assessments. Lexical variation was type/token ratio, which could be influenced by text length.

A different definition of words also influenced type/token values. The lexical variation could not distinguish word quality and only reflected the different words used in a text. They also mentioned that the less frequently used measures, like semantic variation (Mendelsohn, 1981), lexical quality (LQ) (Arnaud, 1992), T-unit and error-free T-unit, were problematical.

Laufer and Nation (1995) introduced the Lexical Frequency Profile (LFP) calculation process, claiming it overcame the shortcomings inherent in earlier measurements. The LFP presented the percentage of words at different frequency levels in learners' essays. One essay could be divided into the first 1000 words, the second 2,000 words, UWL (University Word List), and not-in-the-list words. The calculation could be done through a computer program, VocabProfile, and the word unit definition of the program was in terms of word tokens, word types, and word families. The word family calculation was every word family distinction at level three, described by Bauer and Nation (1993).

To validate the reliability and validity of the LFP as a measure of lexical richness, Laufer and Nation proposed three aims in their research. First, the profile results would not be influenced by changing topics and would remain stable across the same research subjects. Second, finding the correlations between the existing vocabulary measures in active use or receptive was also an effective way to validate the LFP concurrently. Laufer and Nation adopted the Vocabulary Levels Test's active version (Nation, 1983) in their studies. Third,

since lexical richness measures were a part of language proficiency levels, they wanted to see if the LFP could distinguish between different groups. To resolve the reliability and validity of the LFP, they put forward two research questions for each side. The first two questions aimed at the validity aspect. Q1 was whether significant differences in the LFP existed between learners' proficiency levels, and Q2 was to investigate if the LFP correlated with the active version of the Vocabulary Levels Test. Q3 and Q4 were intended for reliability. Q3 explored the LFP correlations between two essays written by the same learners. Q4 concerned the correlation between the percentage scores at different frequency levels of the two writings produced by the same students.

Three groups of learners of different proficiency levels were included in their paper. 22 EFL learners were students at Victoria University, New Zealand, and their English levels were assumed to be low intermediate. Twenty subjects were undergraduates from Israel, and the learners were in their first semester in the English Language and Literature department. The remaining group was 23 learner undergraduates from the same background as the second group of students, but the only difference was that they had finished their second semester. Each subject wrote two compositions within one week during the data collection process during their class time at 300–350 words. All the topics were very general, and their writing was counted as part of their final grades. For the data processing, Laufer and Nation chose the first 300 words of each writing for calculation, and each running word was treated as a word family. In their research, four values could be obtained for each essay: the first 1,000 words, the second 1,000, the University Word List (UWL), and the not-in-the-lists words.

In response to their first research question, their results showed that the proportion of first-1000 frequent word families significantly differed among the three groups of learners. Group 1 used the first 1,000 frequent words more than group two and group three in two compositions. Regarding the second 1,000 frequent terms, there were no significant

differences between the three groups in the two writings. As for the UWL, in the first essay, the lower proficiency level students in group one used fewer UWL words than the other two groups. However, significant differences only existed in their second essay among the three groups. It was apparent that higher-level students used more not-in-list words. Laufer and Nation proposed these results indicated that students with rich vocabulary would have more language knowledge, proving the validity of the LFP in revealing these differences.

To respond to their second research question, Laufer and Nation adopted the active version of the Levels Test (Nation, 1983). Since there were no first-1000 frequent words or 'not-in-lists' words in the Active Levels Test, Laufer and Nation combined the answers in the Active Levels Test. Their results showed that students who got high marks on the Active Levels Test would have a high score on both UWL and 'not-in-lists' words. Negative correlations were reported between the first-1000 words, and no correlations were reported between the second-1000 words and the Active Levels Test. Laufer and Nation's research questions three and four were about the reliability of their measure, and their results showed that group one and group two appeared stable with the two essays. However, for the high proficiency students (group three), there were differences in the first 1,000, UWL and not-in-lists words, indicating higher-level students tended to produce more words across different writing topics.

Thus, Laufer and Nation concluded that the LFP was a valid and reliable tool in assessing lexical use in writings, which had been proved to remain stable within two essays by the same participants and could discriminate between different levels. The LFP also correlated well with a lexical use measure. Using computer programs to deal with essays was an effective tool in research. They also emphasised that learners' productive vocabulary in writing could reflect learners' vocabulary size. They asserted it was crucial to increase the possibilities of using vocabulary knowledge and adjusting it to a teaching program.

Critique

Laufer and Nation's (1995) research influenced lexical analysis studies. They argued that the LFP measurement showed the proportion of words at different frequencies in a way superior to other lexical richness measures explaining productive vocabulary use. In their research, the LFP effectively reflected vocabulary size in use and distinguished students' proficiency levels and correlated well with the active version of the Vocabulary Levels Test (Nation, 1983). Despite all these strengths, the LFP still has some weaknesses. In the following critique, I address five concerns. These concerns are (i) the use of vocabulary lists, (ii) the concurrent validity of the Lexical Frequency Profile (LFP) when using 1000-word frequency bands, (iii) the assumption that lexical richness is equivalent to frequency, (iv) the selection of the first 300 words for analysis, and (v) the use of word families as a unit for word counting.

The first potential problem relates to the vocabulary lists used in their studies. Laufer and Nation (1995) divided the vocabulary lists into four scales: the first 1,000 frequency words; the second 1000 frequency words; the University Word List (containing the academic word list); and the not-in-lists words. All the word lists in their studies are based on Nation's (1983) assertion that initially word lists were derived from the General Service List (West, 1953), which are out-of-date in terms of current corpus development research. Though the word lists have been updated based on the British National Corpus (BNC Consortium, 2007) and the Corpus of Contemporary American English (COCA; Davies, 2010), some words are still classified into the not-in-lists frequency. This allows researchers to quickly identify the percentage of not-in-lists terms that make up a relatively large proportion of learners' writings as a result of the incomprehensibility of vocabulary lists. The highest-level

participants in group three showed the not-in-lists values of 7.5% and 8.7%, respectively, in their first and second writings.

Second, to address the concurrent validity of the LFP, Laufer and Nation (1995) adopted the active use of the Vocabulary Levels Test (Nation, 1983) to determine the correlations between the LFP and the active version of VLT. Their results showed no correlations between the second 1,000 words and the LFP. They argued that low- and high-vocabulary-size students have used the 'middle level' words (the second 1,000 words), which shows that the second 1,000 words cannot distinguish students' vocabulary size, implying that words at the second 1,000 level need further investigation. Two assumptions undermine these arguments. One is that the eighteen sentences in their active use of VLT cannot represent students' vocabulary knowledge of their second 1,000 words level of the LFP. The other is that the hasty classification of the second 1000 words is questionable. Meanwhile, Kremmel (2016) commented that 'the traditional 1,000-item frequency bands are not optimal' (p. 976). He indicated that 'frequency is a continuum' and 'vocabulary test developers have taken frequency division as a tradition, arguably for the sake of being able to work with round numbers' (p. 980) despite the lack of empirical evidence. Kremmel also argued that a 500-lemma-based frequency division was a more fine-grained band than a 1000-lemma-based one.

Third, frequency is widely used in evaluating learners' vocabulary knowledge. Laufer and Nation explained that lexical richness equates to frequency in their research, which is problematic considering modern computation measures. Read (2000, p. 200) mentioned that lexical richness includes four components: type-token ratio or lexical variation; lexical sophistication; lexical density; and the number of errors. Lexical variation means lexical diversity measuring the number of unique words in writing and speaking contexts. Malvern et al. (2004) similarly queried the assumption of 'lexical diversity and lexical sophistication as

being subsumed under vocabulary richness' (p. 5). Likewise, Treffers-Daller et al.'s (2018) paper showed that lexical diversity scores show significant correlations with vocabulary knowledge scores, writing levels, and language proficiency. Considering the more recently devised measures, using frequency to measure lexical richness for written text appears ineffective.

Fourth, in Laufer and Nation's study, they chose the first 300 words for analysis for reasons that they did not fully explain. However, we can assume that the LFP is also sensitive to text length, like the type-token ratio. Thus, whether the first 300 words offer the most appropriate word selection needs further validation.

Fifth, another issue in Laufer and Nation's research pertains to how to deal with the various levels of word families. They use the word family as a unit to treat words. McLean (2018) proposed that the flemma is the most appropriate word counting unit for language learners in assessing their word knowledge level. Kyle (2019) indicated that many studies appeared to lemmatise texts; however, they were actually flemmatising them because the main difference between a flemma and a lemma is that a lemma is sensitive to the part of speech while a flemma is not. The inflected and derived forms in word families reflect learners' knowledge of the language. Using the word family as a counting unit is more appropriate for high-level learners who have obtained knowledge of word families. In contrast, lower-level learners may lack the word family knowledge. For this reason, distinguishing the level of the word family used to reflect learners' knowledge of words is significant. In Treffers-Daller et al.'s (2018) paper, they deployed three kinds of word counting units in computing learners' essays: types, lemma, and word families (up to level 3). Their research has shown that type is the most effective unit in predicting proficiency levels.

In conclusion, Laufer and Nation's (1995) paper is significant for exploring the relationship between vocabulary knowledge and writing proficiency. They pioneered the

active version of the Levels Test, later known as the Productive Vocabulary Levels Test (PVLT), within the research community. Their findings show a significant correlation between vocabulary knowledge and the LFP. However, their research has shortcomings, particularly regarding using the LFP to evaluate written texts. The primary issues are (i) the use of outdated word lists that result in a high percentage of not-in-the-list words in writing; (ii) the LFP's frequency-based method for validating lexical richness in writing; (iii) the inappropriate treatment of lexical richness as equivalent to frequency, especially considering more recently devised measures such as lexical diversity; (iv) the choice of the first 300 words for the writing samples; and (v) the use of the word family as a word counting unit. As a result, Laufer and Nation's innovative method, the LFP, needs to be used judiciously in future studies aimed at assessing lexical richness.

2.2.2 Laufer, B., & Nation, P. (1999): A Vocabulary-Size Test of Controlled Productive Ability.

Laufer and Nation's (1999) paper introduced a reliable and valid measure for testing productive vocabulary knowledge. The measure consisted of five frequency levels: 2000, 3000, 5000, UWL, and 10000, and it proved effective in distinguishing proficiency bands. The authors aimed to enhance the effectiveness of vocabulary testing through a controlled productive vocabulary measure. The Vocabulary Levels Test (VLT) developed by Nation (1983, 1990) and Meara's Eurocentres Vocabulary Size Test (Meara & Buxton, 1987) are both convenient to administer during class time and can evaluate numerous words at once. As different types of vocabulary prioritise varying degrees of vocabulary knowledge (Paul et al., 1990), Richards (1976) and Nation (1990) have proposed multiple scales of word knowledge. Nation has also emphasised the importance of including both receptive and productive measurements of multidimensional vocabulary knowledge to understand learners' vocabulary

knowledge comprehensively. Therefore, it is essential to incorporate a variety of vocabulary assessments to evaluate learners' vocabulary knowledge accurately.

Laufer and Nation (1995) examined lexical richness in writing by utilising word frequency, while Laufer and Nation (1999) utilised different frequency levels in the Vocabulary Levels Test (VLT) to assess language learners' vocabulary size. Nation explained the rationale behind using frequency levels to evaluate vocabulary knowledge. He noted significant differences existed between the frequency of word use, with the first 1000 words accounting for about 75% of written and 84% of spoken language use. Conversely, the English language also contains many words (Goulden et al., 1990) that are seldom used. Therefore, it is crucial to focus on which words are worth attention. The distinction between high and low-frequency words has significant implications and can enable teachers to access their students' vocabulary knowledge and provide valuable feedback on their vocabulary development.

In Laufer and Nation's (1999) paper, they noted learners tend to avoid using infrequent words when assigned writing tasks by their teachers but will use them more freely when writing independently. They stated that this reluctance could indicate a lack of confidence in their word knowledge. Their earlier paper (Laufer & Nation, 1995) had focused on learners' voluntary use of vocabulary knowledge, as measured by the Lexical Frequency Profile (LFP). In contrast, controlled productive vocabulary measures tend to focus on the ability to use words under the pressure of teachers or researchers, whether in a free-writing context or a more constrained setting such as a sentence completion task. The controlled productive task in their study followed the latter format of a sentence completion task. Laufer and Nation used sentence context to elicit target words, providing cues with the first few letters to remove ambiguity. The task comprises eighteen sentences selected from each word

frequency level: 2000, 3000, 5000, University Word List (UWL), and 10000. An example from their task is as follows:

The book covers a series of isolated epi____ from history. (Laufer & Nation, 1999, p. 37)

Laufer and Nation (1999) conducted two studies in their paper. The first study examined the validity of the controlled productive task; the second study aimed to check the consistency of four parallel versions of the controlled productive task. To verify if the task could differentiate among proficiency levels, their study included four sets of learners: high school 10th graders (n=24); 11th graders (n=23); 12th graders (n=18); and university students studying English (n=14). They gave each student a controlled productive vocabulary test, and minor spelling and grammatical mistakes were disregarded. During the test process, three L1 English speakers offered the students help to retrieve the vocabulary by giving modified sentences context or adding one more letter for the target vocabulary. Students were allocated six scores for the corrected vocabulary and the retrieved ones. They were awarded one point for each correct response on each frequency level, and the final points were the total correct responses across all five frequency levels.

The results in their first study showed that all participants had an internal consistency of 0.86 using the Kuder-Richardson KR21 formula. ANOVA analysis with Duncan post-hoc was used for each vocabulary frequency level and the overall vocabulary scores. As Table 2.1 shows, participants' mean scores on each vocabulary level increase along with the proficiency levels. Their total scores increase from 21.7 points in 10th grade to 55.8 points for the university-level participants. The results in Table 2.1 suggest that participants' vocabulary knowledge decreases as word frequency decreases, which applies to participants at all four different language levels.

Table 2.1

Mean Scores and F-Tests for Four Proficiency Level Groups on the Five Levels and the Total Score of the Original Productive Levels Test

	10th grade (n=24)	11th grade (n=23)	12th grade (n=18)	University (n=14)	F-test
2000 level	11.8	15	16.2	17	17.9 p=.0001
3000 level	6.3	9.3	10.8	14.9	21.2 p=.0001
UWL level	2.6	5.3	7.4	12.6	34.6 p=.0001
5000 level	1.0	3.9	4.7	7.4	12.6 p=.0001
10000 level	0.0	0.0	0.9	3.8	13.6 p=.0001
Total	21.7	33.4	40.1	55.8	32.6 p=.0001

Note. Adapted from "A vocabulary-size test of controlled productive ability," by B. Laufer and P. Nation, 1999, *Language Testing*, *16*(1), p. 39. (https://doi.org/10.1177/026553229901600103)

In their second study, Laufer and Nation's paper conducted three parallel versions made by Norbert Schmitt to compare with the first version of the productive vocabulary levels test. They called the first version A, and the three parallel versions B, C and D. Across these four versions, there were different vocabulary items for each frequency level. They used different participants in the second study from their first study. The test contents were also nonidentical. In their second study, each participant took the four parallel versions for each frequency level, not the whole test, and the 10000-frequency level was not used for the second study because the EFL learners did not have a good knowledge of this level. Table 2.2 shows the correlation results among the four versions of the PVLT across the first four frequency levels. Significant correlations exist between the first three frequency levels across four different versions. Regarding the 5000 level, strong significant correlations only exist between version A and version C, whereas no significant correlations exist among version A, version B, and version D.

Table 2.2

Correlations Between Four Versions of the Productive Vocabulary Levels Test at Four of the Five Frequency Levels in the Tests

	A/B	A/C	A/D	B/C	B/D	C/D
2000 level (n = 45)	.82*	.82*	.78*	.83*	.81*	.77*
3000 level (n = 36)	.71*	.70*	.82*	.82*	.71*	.80*
UWL level (n = 33)	.75*	.80*	.84*	.83*	.76*	.80*
5000 level (n = 18)	.72 (p = .004)	.83*	.69 (p = .003)	.49 (p = .1)	.77 (p = .003)	.67 (p = .006)

Note. *Significant at .0001 level. Adapted from "A vocabulary-size test of controlled productive ability," by B. Laufer and P. Nation, 1999, *Language Testing*, *16*(1), p. 43. (https://doi.org/10.1177/026553229901600103)

Table 2.3 shows that the two parallel productive vocabulary level tests correlate well at the first four frequency levels. Strong significant correlations exist between each frequency level across different versions. Their paper suggested that all four versions can be used for diagnostic purposes, and two of the newly created versions (version C and version D) are recommended for test and retest purposes.

Table 2.3

Two Equivalent Forms With Similar Means and a Good Correlation at Each Level

Level	2000 B/C	3000 C/D	5000 A/C	UWL C/D
Means	6.7/6.3	3.8/3.9	3.7/3.5	5.1/5.7
Standard deviations	3.3/3.3	2.3/2.6	2.3/1.7	2.9/3.8
Correlations	.83	.80	.82	.80

Note. Adapted from "A vocabulary-size test of controlled productive ability," by B. Laufer and P. Nation, 1999, *Language Testing*, *16*(1), p. 44. (https://doi.org/10.1177/026553229901600103)

Considering the results shown in Table 2.2 and Table 2.3, Laufer and Nation concluded that the Productive Vocabulary Levels Test (PVLT) is a valid tool for measuring vocabulary development with different versions and is also easy to manage. Learners can finish the test within a short time. In addition, the marking process is straightforward to handle because there is only one correct answer for each sentence. They highlighted that future research could investigate more issues using the PVLT with its receptive version and the LFP (Laufer & Nation, 1995; reviewed in section 2.2.1).

Critique

Laufer and Nation's study validated the PVLT task across four different versions for participants with four different proficiency levels. They stated the PVLT is a powerful supplement to receptive vocabulary measures like the Vocabulary Levels Test (Nation, 1983; 1990) as it can 'look more effectively at the breadth of vocabulary knowledge' (Laufer & Nation, 1999, p. 45). They also argued that the PVLT is a reliable tool for assessing vocabulary growth by comparing learners' performance at different frequency levels. Despite all these strengths, the PVLT still has some weaknesses. In the following critique, I address two such potential concerns. These concerns are (i) using 18 words of each frequency level and (ii) the pre-determined words for the PVLT.

First, the problem relates to using 18 items in each frequency band to represent participants' vocabulary knowledge of 1000 words. Laufer and Nation's description of the percentage score at a frequency level can be interpreted as the indicator of the number of mastered words at that level. For instance, if a learner knows nine words among 18 in the UWL, it represents a learner with 50% knowledge of the UWL, meaning the learner knows 418 out of 836 words. This approach seems inaccurate in predicting learners' knowledge of a thousand words only based on 18 items, and has been criticised by Meara and Fitzpatrick (2000) in connection with higher-level learners who have more knowledge of infrequent

vocabulary but are still not able to complete all the preset words. They may know other infrequent words and use them as substitutes for the fixed ones. In contrast, we cannot say a learner has acquired all the vocabulary knowledge at a certain frequency level just by correctly filling in 18 random items. Suppose the selected sentences and the cues can exactly elicit learners' word knowledge of all the 18 items. In that case, the learner only has very limited knowledge to that level, which may well usually be indicative of low-level learners. Specifically, using only 18 words to represent learners' knowledge about the 5,000 and 10,000 words is also problematic because words at the 5,000 and especially 10,000 frequency levels, are infrequent for English language learners. We cannot confidently extrapolate that participants who can fill out all 18 words in this frequency band have knowledge of all the words at this level.

Second, the PVLT task is limited when it comes to the pre-determined words that are to be filled in in the sentences. This reduces exposure opportunities for participants who may know other words not included in the pre-determined list of 18 words for each frequency level. The PVLT task may not accurately measure participants' vocabulary knowledge for a given frequency level that includes 1,000 words. This is because it only requires pre-determined words, but participants may fill in other synonymous words besides the given 18 words. Laufer and Nation (1999) addressed this issue by providing participants with one additional letter to help them produce the target words. However, this method is not practical when testing numerous participants simultaneously.

There are instances when participants may fill in a semantically and grammatically correct word using the first few given letters. However, it may be a different word altogether. Laufer and Nation (1999) did not explain how to address such cases during testing. These problems are due to the limited exposure opportunities presented by the PVLT task.

Laufer and Nation's (1999) study is undoubtedly important because it proposed a method for testing vocabulary knowledge and created four parallel versions for further research. However, there are still potential issues with the PVLT task: (i) the first problem is related to using only 18 items to represent each frequency level, especially for the 5k and 10k levels; (ii) the second limitation is related to the fact that the PVLT task requires participants to fill in pre-determined words.

2.2.3 Meara, P., & Fitzpatrick, T. (2000): Lex30: An Improved Method of Assessing Productive Vocabulary in an L2.

Meara and Fitzpatrick (2000) introduced a productive vocabulary measurement tool, Lex30. Meara and Fitzpatrick's paper sought to build an effective way to assess the productive vocabulary of second language learners (L2). They based their examination on the understanding that language learners with a larger vocabulary size will have higher language proficiency (Meara & Jones, 1988).

An important issue put forward by Meara and Fitzpatrick was that assessing productive vocabulary appears much more difficult than assessing receptive vocabulary knowledge. The major reason for this is that the productive aspects of writing and speaking are so context-specific, and we cannot infer the true L2 vocabulary size from limited productions. Therefore, inventing simple tasks to activate large vocabulary quantities is also challenging.

Meara and Fitzpatrick (2000) wrote that extant productive vocabulary tests created by Laufer and Nation (1995; 1999) are problematic. They indicated that the controlled productive vocabulary tests (Nation, 1983; Laufer & Nation, 1999) prompt learners to produce preset words by offering learners a sentence context and a few beginning letters of the target words, and learners are then required to complete the missing letters. For example:

The book covers a series of isolated epis_____ from history.

The problem with controlled productive vocabulary tests is that they work mainly for lower-level students with a limited vocabulary size which can cover a high proportion of these tested words. Testing the actual vocabulary size through 18 target words among five frequency bands (2000, 3000, 5000, University Word list and 10,000) is also very hard. Language learners only need to fill in the exact words based on the given English letters, which means that other infrequent words that learners may know will not be tested. The free productive vocabulary tests, like the LFP (Laufer & Nation, 1995), give learners a written or spoken topic and then utilise lexical frequency to describe the quality of their production of vocabulary knowledge, which means learners who can produce a higher percentage of infrequent words have a higher productive vocabulary knowledge. Meara and Fitzpatrick observed that the Laufer and Nation (1995) tests (the LFP) are context-limited, even though a general topic will be selected. We cannot determine whether the elicited words in their compositions truly reflect learners' productive vocabulary knowledge. Meanwhile, they also mentioned that Laufer and Nation's (1995) test is ineffective, because most writing contains collections of high-frequency words. Moreover, writing samples (which usually need to be 300 English words) also take at least two hours of class time, and it is hard for second language learners to complete.

Considering the practical problems highlighted above, Meara and Fitzpatrick sought more efficient ways to elicit productive vocabulary data from language learners. They, therefore, developed Lex30, a task based on word association, using thirty cue words. They required test-takers to write down at least three responses; the maximum number of responses is 120 (30 cues multiplied by 4). Lex30 is like a free productive vocabulary task, but test-takers do not need to write strictly preset target words, so Lex30 will elicit more varied words while ultimately being less constrained by contextual concerns. As their paper described, all

stimulus words in Lex30 meet the following three criteria. First, they are highly frequent stimulus words chosen from Nation's (1984) first 1000-word lists. Second, they ensure the stimulus words can generate a broad range of responses, not just single or dominant ones; for example, they exclude words like *black* or *dog* that elicit only a narrow range of responses. Third, each stimulus word usually generates a range of uncommon response words because at least half of the responses given by L1 participants in Lex30 are beyond Nation's (1984) first 1000-word lists.

Meara and Fitzpatrick's study reported on their first use of Lex30, in which they pilot the task with 46 participants. The participants are 46 adult L2 English speakers, with levels ranging from high-elementary to proficiency and with mixed L1 backgrounds. Test-takers were required to write down the responses (maximum of four) for each stimulus word. They were given 30 seconds for each cue word, and the whole task took 15 minutes. To explore the extent to which the Lex30 task provides an indication of vocabulary knowledge, Meara and Fitzpatrick compared their results alongside a test of a yes/no format, a receptive vocabulary knowledge measure. Each participant took the yes/no test within the same week (Meara & Jones, 1990). To score the responses generated from Lex30, they discarded the stimulus words, and the responses were lemmatised so that inflected and partially derived forms were eliminated. Their lemmatisation criteria were level 2 and level 3, as outlined (see Table 2.4) by Bauer and Nation (1993). They dealt with each response by a frequency program (similar to Heatley & Nation, 1998), and every word was classified into its corresponding frequency level. In this program, level 0 means high-frequency structure words, proper names and numbers, and level 1 words are the first 1000 frequent content words in English. The responses within level 0 and level 1 got zero points, while any responses beyond these two levels were given 1 point for each word.

Table 2.4

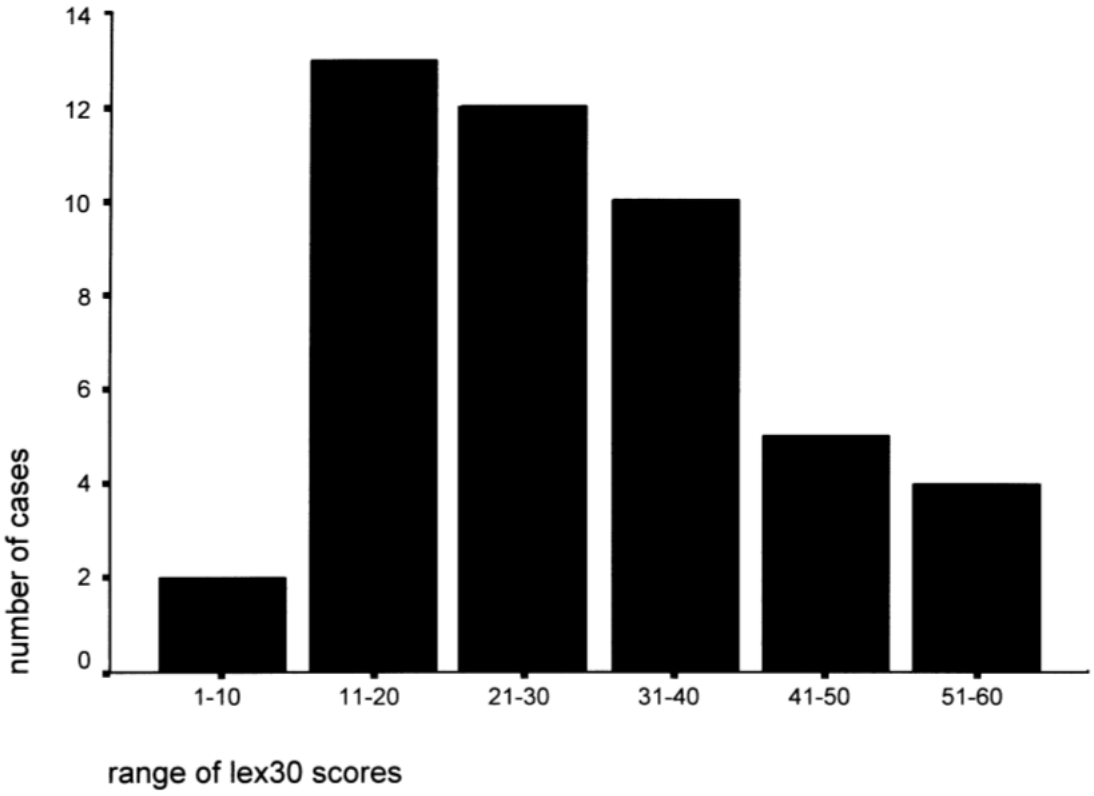
Lemmatisation Criteria of Level 2 and Level 3

Level 2: Inflectional suffixes	Level 3: Most frequent and regular derivational affixes
* plural	* -able (does not apply to nouns)
* 3rd person singular present tense	* -er
* past tense	* -ish
* past participle	* -less
* -ing	* -ly
* comparative	* -ness
* superlative	* -th cardinal - ordinal only
* possessive	* -y adjectives from nouns
	* non-
	* un-

The task results were as follows. Figure 2.1 shows that most words belonged to the first 1000 words, but some test-takers produced large proportions of words beyond level 0 and level 1. Figure 2.2 shows the relationship between Lex30 and the yes/no test. The correlation between the two tests was 0.841 ($p<0.01$). The paper also pointed out that if we look closely at the results in Figure 2.2, we can find that some test-takers, whose scores lie above the line, had a higher productive vocabulary knowledge than their receptive vocabulary knowledge as indicated by their yes/no test scores. In contrast, scores below the line have higher yes/no scores than for their productive vocabulary. This figure also suggests that test-takers with higher yes/no scores also achieved higher Lex30 scores.

Figure 2.1

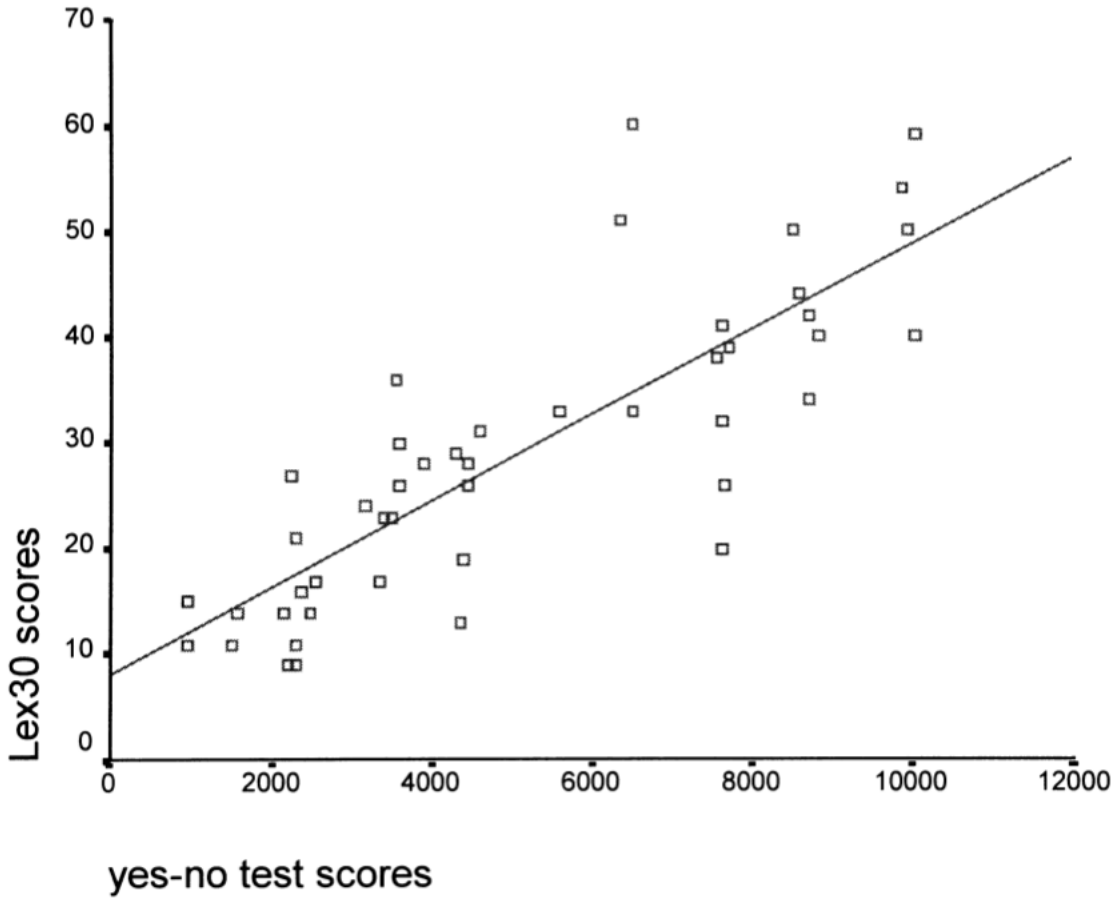
Distribution of Lex30 Scores



Note. Adapted from "Lex30: An improved method of assessing productive vocabulary in an L2," by P. Meara and T. Fitzpatrick, 2000, *System*, *28*(1), p. 25. (https://doi.org/10.1016/S0346-251X(99)00058-5)

Figure 2.2

Comparison of Yes/No Test Scores and Lex30 Scores



Note. Adapted from "Lex30: An improved method of assessing productive vocabulary in an L2," by P. Meara and T. Fitzpatrick, 2000, *System*, *28*(1), p. 25. (https://doi.org/10.1016/S0346-251X(99)00058-5)

In short, Meara and Fitzpatrick's (2000) study examined the performance of Lex30 as a productive vocabulary task. The high correlations between the infrequent (equal to or over 2000 frequency words) words and the yes/no test (a receptive measure) can support the concurrent validity between Lex30 and the yes/no test. They emphasised that the advantages

of Lex30 are that it gives the students every chance to obtain scores no matter what words they produce and is quite easy to handle being less time-consuming and demanding less effort both of teachers and students. Their paper also suggested that language learners may experience undesirable vocabulary knowledge development while acquiring vocabulary knowledge. Lex30 can also be a diagnostic tool to pinpoint weaknesses and design training programs for vocabulary knowledge development.

Critique

The productive vocabulary task, Lex30, was devised and validated in Meara and Fitzpatrick's (2000) study. Many advantages have been proposed, which include: the practical functions in use; the deliberate selection of the stimulus words; the high correlations (0.841, $P<0.01$) with a receptive vocabulary measurement (a yes/no format test); generating words in a relatively less constrained way; and the capacity to be developed as a diagnostic tool for specific people. Despite these strengths, I should mention some problems with Meara and Fitzpatrick's study. In the following critique, I address one primary concern: Not all the elicited words in the Lex30 task can accurately represent participants' productive vocabulary knowledge.

The concern is that the words generated in the Lex30 task cannot represent language learners' productive vocabulary knowledge, especially with low-level learners.
Since Lex30 only requires test-takers to spell out and write the elicited words correctly, it can generate many words from them. Learners can write many elicited words, but they may not really know how to use them in diverse contexts. The recall definition means test-takers are presented with some stimulus to elicit the words from their memory (Read, 2000). From this definition, Lex30 is a task, to some extent, directly relating to the recall process (Fitzpatrick & Meara, 2004). Since the correct recall of the targeted words also works for vocabulary use

(Read, 2000), it is very hard to distinguish which percentage of words are just recalled and which kind of recalled words the learners can actually use correctly semantically. Learners who can write the words recalled from their memory may not be able to use them correctly when faced with situations involving more context, such as words with more complex semantics, collocations, or requiring grammatical knowledge. Walters (2012) (reviewed in the following section) has tried to distinguish between recall and productive use through a sentence elicitation task combined with a depth of vocabulary knowledge method developed by Wesche and Paribakht (1996).

Considering this problem, of whether Lex30 can be a validated tool, it needs to be further validated, because as Meara and Fitzpatrick (2000) concluded, there are still 'a number of outstanding issues concerning the reliability and validity of the Lex30 methodology' (p. 28). Accordingly, Fitzpatrick and Clenton (2017) further validated Lex30 by comparing it with other productive vocabulary knowledge tasks, as will be reviewed fully in section 2.2.5.

2.2.4 Walters, J. (2012): Aspects of Validity of a Test of Productive Vocabulary: Lex30.

Walters's (2012) study examined the construct validity of Lex30. The concurrent validity was investigated with two productive vocabulary tasks: the Productive Vocabulary Levels Test (PVLT; Laufer & Nation, 1999) and a translation test. Meanwhile, the use or recall issue was also considered in evaluating productive vocabulary knowledge. Moreover, to differentiate the proficiency levels of the L2 language learners, Walters's paper includes three groups of subjects. The results indicated that Lex30 is a convincing measure in assessing L2 productive vocabulary knowledge, but whether it measures use or recall depends on the proficiency levels of the particular L2 learners.

Walters's (2012) study first introduced the background of the previous Lex30 research. For Walters, productive vocabulary testing methods were divided into two approaches: examining the vocabulary knowledge at various frequency bands like the PVLT (Laufer & Nation, 1999, reviewed in section 2.2.2) or extracting more words from the test-takers and then dividing the target words into frequency, such as the Lexical Frequency Profile (LFP; Laufer & Nation, 1995, reviewed in section 2.2.1) and P_Lex (Meara & Bell, 2001). As explained above, the PVLT gives the first several letters; test-takers are asked to fill out the target word in sentence context. Due to the word restrictions, the PVLT is classified as a controlled productive vocabulary task. The LFP is an essay writing task, and the data were computed by the website Vocabprofile. The elicited writings are presented by the percentage of words of each frequency, namely the K1 (first 1000 words); K2 (the second 1000 words); the Academic Word List (AWL; Coxhead, 2000); and the off-list words. P_Lex (Meara & Bell, 2001) is also a free writing task, but it utilises a different method (the lambda score) to count the infrequent words in each segment. It is claimed that P_Lex is more applicable to lower-level L2 learners. Considering the time-consuming difficulties and high percentage of frequency words in free writing, Walters's paper uses Lex30, a task based on word association. Participants must write the first four words that immediately come to mind to reply to the stimulus word. After lemmatising all the responses, the frequency levels lists process the answers.

Walters's paper mainly reports on four experiments in which she examines the validity and reliability of Lex30. Meara and Fitzpatrick (2000), reviewed in section 2.2.3, was the first study to develop a Lex30 task, with 46 EFL adult learners showing the strong correlations of 0.841 ($p<.01$) between the receptive vocabulary measure (a yes/no test) and Lex30. However, the validity and reliability needed to be further investigated. Thus, Fitzpatrick and Meara (2004) sought to further validate the Lex30 task through three groups

of participants. A test-retest validity was distributed twice with the same sixteen subjects, with a gap of three days between tests. Correlation scores between the two tests were relatively high (.866, $p<.01$), and the participants could produce new words but at the same frequency within two different Lex30 versions. They also looked at the validity with another 46 L1 English speakers by comparing them with those L2 learners whom Meara and Fitzpatrick (2000) had tested. The results showed that L1 English speakers have a higher lexical score than L2 learners in general, while some L2 test-takers have higher vocabulary scores than L1 English speakers. A further step was taken to examine high-level L2 learners. Fitzpatrick and Meara (2004) concluded that Lex30 operates well in distinguishing different proficiency levels of L2 English speakers. Fitzpatrick and Meara studied the concurrent validity of Lex30 by comparing it with the PVLT and a translation test. Their results showed moderate to strong correlations among the three tests: the correlations between Lex30 and the PVLT fall at .504 ($p<0.01$); the correlations between Lex30 and the translation test fall at .651 ($p<0.01$); and the correlations between the PVLT and translation test fall at .843 ($p<0.01$). Thus, Fitzpatrick and Meara's (2004) paper explained that these tests tap different aspects of word knowledge, as in Nation's (1990) description. Fitzpatrick and Meara (2004) further noted that Lex30 could elicit a representative vocabulary set, however it also achieved simple vocabulary recall in accordance with Read's (2000) definition. In addition, Fitzpatrick and Clenton's (2010) paper further analysed the validity and reliability of Lex30 in terms of internal validity, its reliability in reflecting vocabulary improvement, and aspects of construct validity. Building on the above studies, Walters (2012) conducted her study to further validate the Lex30 task.

Walters used 87 L2 English learners divided into three groups based on their English language experience, and all the participants had the same background (L1 Turkish). A

detailed description of the participants can be seen in Table 2.5. She used four data collection tools: Lex30; the PVLT; a translation test; and a sentence elicitation task.

Table 2.1

Proficiency Level Descriptions

Participant Group	English Language Experience	Proposed Proficiency Level
Bilkent University group (N = 32)	Currently studying in an English-medium MATEFL program; a minimum of 2 years of English language teaching experience.	Advanced
Erciyes University group (N = 25)	Completed 1-year English preparatory program at university; currently studying in an undergraduate level English Language Teaching program, 3rd year.	Intermediate
Hacettepe University group (N = 30)	Currently studying in a 1-year English preparatory program at university, in second semester.	High beginning

Note. Adapted from "Aspects of validity of a test of productive vocabulary: Lex30," by J. Walters, 2012, *Language Assessment Quarterly*, *9*(2), p. 176. (https://doi.org/10.1080/15434303.2011.625579) MATEFL=Master of Arts program in teaching English as a foreign language.

Walters presented three issues in her results and discussion section: (i) the ability to differentiate among proficiency levels; (ii) concurrent validity; and (iii) the recall or use problem. Table 2.6 shows the Lex30 results for the three different proficiency levels' students. Moreover, the means of the three groups are different by ANOVA analysis ($p<.001$). the Post-hoc Scheffé value ($p<.01$) indicated that the three groups were significantly different, but some overlaps exist among these groups. The results show Lex30 can distinguish different proficiency levels, but the overlaps remind us that the students who are at the same level still differ in their Lex30 scores to some extent.

Table 2.2

Lex30 Task Results

	No.	*Minimum*	*Maximum*	*M*	*SD*
High beginning	30	16	37	27.23	5.722
Intermediate	25	20	59	36.72	10.048
Advanced	32	28	77	55.84	11.706
Whole group	87	16	77	40.48	15.549

Note. Adapted from "Aspects of validity of a test of productive vocabulary: Lex30," by J. Walters, 2012, *Language Assessment Quarterly*, *9*(2), p. 179. (https://doi.org/10.1080/15434303.2011.625579)

Regarding concurrent validity, Walters' (2012) paper manipulated both the PVLT (only in 2,000 and 3,000 levels) and a translation test choosing 60 words separately from the 1,000, 2,000 and 3,000 levels with the Brown Corpus (Kučera & Francis, 1967). Students were asked to translate the Turkish version of all 60 words into English according to the first letter of the words given. Looking at the findings in Table 2.7, strong correlations exist between the PVLT and the Translation Test (r=.936), and the reasons may be that they both use 2,000 and 3,000 frequency levels. Significant correlations exist between Lex30 and the PVLT (r=.772) and between Lex30 and the Translation Test (r=.745). The strongest and the most significant correlations exist between the PVLT and the Translation Test, followed by strong significant correlations between Lex30 and the PVLT, and then the correlations between Lex30 and the Translation Test. Lex30 shows positive correlations with two productive vocabulary tests (the PVLT and the Translation Test), and Lex30 responses can predict learners' performance on these two tests relating to productive vocabulary knowledge to some extent.

Table 2.3

Correlations, Lex30, the PVLT, and Translation Test

	PVLT	Translation Test
Lex30	.772 ($p < .001$)	.745 ($p < .001$)
PVLT		.936 ($p < .001$)

Note. Adapted from "Aspects of validity of a test of productive vocabulary: Lex30," by J. Walters, 2012, *Language Assessment Quarterly*, *9*(2), p. 181. (https://doi.org/10.1080/15434303.2011.625579)

To validate the construct validity of Lex30 regarding the recall or use problem, Walters' paper used a sentence elicitation task to investigate whether students can use the infrequent words falling at Level 3 (AWL and off-list words) in the Lex30 task. A scoring rubric estimated target words representing students' general vocabulary ability. During the scoring process, she did not penalise grammatical errors. The results showed that higher-level students can produce more sentences than intermediate or low-level students. As shown in Table 2.8, the intermediate and advanced students could use words which they produced more properly than high-beginning students. Walters analysed her data using a one-way ANOVA analysis, and the results showed a significant difference in mean scores between the high-beginning students and the two higher-level groups of students (F (2, 77) = 15.628, $p < .001$, $\omega = .94$). However, no significant differences existed between the intermediate and advanced groups. Her results showed that higher-level students could use more correct words produced in the Lex30 task, which shows that Lex30 is a reasonable tool for evidencing productive vocabulary in *use* for higher-level students. Meanwhile, Lex30 can elicit more *recall* words for lower proficiency participants.

Table 2.4

Results of Sentence Elicitation Task

	No.	Minimum	Maximum	M	SD
High beginning	26	.00	100.00	62.7603	27.80417
Intermediate	24	53.85	100.00	81.2868	12.34675
Advanced	30	74.29	100.00	88.8294	6.8171
Whole group	80	.00	100.00	78.0942	20.74436

Note. Adapted from "Aspects of validity of a test of productive vocabulary: Lex30," by J. Walters, 2012, *Language Assessment Quarterly*, *9*(2), p. 182. (https://doi.org/10.1080/15434303.2011.625579)

Critique

Walters' research investigated the validity aspect of the productive vocabulary knowledge measure, Lex30. Her paper has explored three main questions: (i) using Lex30 results to distinguish different proficiency levels; (ii) combining Lex30 with the PVLT and a translation task in evaluating the concurrent validity; and (iii) whether Lex30 works for vocabulary recall/use through a sentence elicitation task. Her study piloted the Lex30 task with vocabulary use or recall aspects, an innovative dimension explored in her paper. Despite these strengths, there are some problems to be discussed with Walters' study. In the following critique, I address these three concerns: (i) The use of part of the PVLT task; (ii) the use of stimulus words for the sentence elicitation task, and (iii) the difficulty in defining the relations between recall/use knowledge and participants' proficiency levels.

First, a potential concern relates to the fact that Walters' paper only uses 2000- and 3000-word frequency levels to assess the concurrent validity issue, excluding other frequency levels, such as the 5000-word level, the University Word Level word list and the 10000-word level. The paper considered the fatigue factor, which may have influenced students' performance during the experimental process while impacting the test results on the PVLT. For the higher-level students, we cannot test the full extent of their productive vocabulary by

using the PVLT because this test paper only offered them the first two frequency levels. The test result of lower-level students may also have been similarly affected. Let us take the following sentence as an example:

We decided to celebrate New Year's E___ together. (Laufer & Nation, 1999; p.47)

This sentence comes from the eighth sentence of a parallel PVLT test of its 5000-word frequency level, and the target word *Eve* is quite easy compared with many target words both in the 2000-word level and 3000-word level, so even the high beginning language learners can fill it in. This caused the high correlations between the PVLT and the translation test, which also chose words from 2,000 to 3,000 frequency levels.

Second, a potential concern relates to the fact that the stimulus words in Walters' paper came from level 3, including AWL and off-list words during the sentence elicitation task. The web-based tool Vocabprofile scored the Lex30 task in Walters' study, and once the file was loaded into the web, it will automatically divide all the words into first 1,000 (K1), second 1,000 (K2), … AWL and off-list words. The off-list words selected by Vocabprofile are problematic; even the names of countries and very simple compound words will be classified into the category of off-list level. These off-list words cannot thus really reflect or be counted as being at an infrequent word level. The website and Walters' paper did not explain or show us the content of the infrequent words and Level 3 words.

Third, another concern in Walters' paper is that the relations between recall/use knowledge and participants' proficiency levels are difficult to distinguish. To explore the recall question Fitzpatrick and Meara (2004) raised, all the words categorised into Level 3 of Lex30 were written by students among three different proficiency levels. Only the sentences marked score 4 (appropriate use of the words in a meaningful sentence) were counted. The results showed a significant difference between the high-beginning and the two higher-level groups but no significant difference between the intermediate and advanced students. Walters

concluded that Lex30 is more valid for higher-level students in productive use and more applicable for vocabulary recall of lower-level students. This means vocabulary recall exists at all proficiency levels, but it is still unclear how much Lex30 can measure the depth (quality) of vocabulary knowledge, so this requires further exploration.

Walters' (2012) study is important because it measured the validity of Lex30 using the PVLT, a translation task, and a sentence elicitation task concurrently. The findings of Walters' study have implications for future research using the Lex30 task, but her study has potential weaknesses: (i) the use of only 2K and 3K frequency levels of the PVLT task for all participants; (ii) the stimulus words for the sentence elicitation task; and (iii) although Walters' paper indicated potential use/recall problems in the Lex30 task, the relations between recall/use vocabulary knowledge in the responses to the Lex30 task and participants' proficiency levels are still not clear.

2.2.5 Fitzpatrick, T. & Clenton, J. (2017): Making Sense of Learner Performance on Tests of Productive Vocabulary Knowledge.

Fitzpatrick and Clenton's (2017) paper investigated the validation of productive vocabulary tests between test-takers' performance and their productive vocabulary knowledge by comparing the Lex30 task (Meara & Fitzpatrick, 2000; reviewed in section 2.2.3) simultaneously with three tests: the LFP (the Lexical Frequency Profile) (Laufer & Nation, 1995; reviewed in section 2.2.1), the BFP (Brainstorm Frequency Profile) and a pilot G_Lex test. As explained above, Lex30 is a task based on a word association format with 30 stimulus English words selected from the 1000 most frequent words, and students need to write the first four words that come into their heads based on the stimulus word. The Lexical Frequency Profile (LFP) was created by Laufer and Nation (1995), in which two compositions of 300–500 English words are required to be written in successive class times.

Then, the writings are processed by the WebVP (www.lextutor.ca), which can divide the vocabulary by frequency. The BFP, a modified LFP task, requires the students to write down as many single English words as possible for the same LFP topics. Fitzpatrick and Clenton's (2017) paper used this task to assess learners' vocabulary knowledge regardless of any grammatical or syntactic restrictions. G_Lex, a gap-fill vocabulary task, uses contextual (sentence) prompts to elicit English words, and it requires test-takers to write the words semantically and syntactically correctly.

To validate the effectiveness of these productive vocabulary knowledge tests and their relationships with test performance precisely, three empirical studies (N=80, 80, 100) were conducted to verify English language learners' performance. Each study begins with an analysis and comparison of two tests and then pioneers the G_Lex test. Study one compared learner performance on Lex30 versus the LFP. Lex30, the word association-based format task created by Meara and Fitzpatrick (2000), is designed to elicit up to 120 English words in total by providing thirty cue words, and has been assessed in many studies (Caton, 2018; Clenton, 2010; Clenton et al., 2020; Fitzpatrick, 2007; Fitzpatrick & Clenton, 2010; Fitzpatrick & Meara, 2004; González & Píriz, 2016; Henriksen and Danelund, 2015; Walters, 2012).

Regarding Fitzpatrick and Clenton's (2017) first study, they compared those two tasks (Lex30 and the LFP). Their paper compared Lex30 and the LFP because they are two different test formats. Even though both tests are for assessing productive vocabulary knowledge and have been investigated in numerous studies, Lex30 is a word association task, whereas the LFP is an essay-writing task. The authors compared the similarities of the two tests considering the vocabulary assessment dimensions, which were put forward by Read (2000, p. 9). Both the LFP and Lex30 are *discrete* and *comprehensive* and rely on *context*. These two tests assess vocabulary knowledge and use it as an independent construct, not an embedded one, and they measure all vocabulary content. As for the *context*, their paper

mentions that it is not easy to distinguish the individual contexts involved in these tests. The participants were 80 (26 female, 54 male) L1 Japanese undergraduates aged between 18 and 21. The researchers used the WebVP to obtain the frequency scores of these tests. The percentage scores of Lex30 and the LFP were calculated because of the different maximum raw task scores: Lex30 can elicit 120 words, whereas the LFP requires students to write 300-350 English words. The results showed that the correlations between Lex30 and the LFP were not significant. One potential explanation beyond the subject knowledge might be how the tasks are scored. Lex30 defines infrequent as being outside the first thousand most frequent word families, whereas the LFP 'infrequency' is adjudged to be outside the 1K and 2K frequent word families (r=.186). Then, the authors adjusted the scoring process; namely, by applying the same frequency definition for Lex30 and the LFP (infrequent = outside 1K band). However, the correlations between Lex30 and the LFP still were not significant (r=.108). Considering the topic, register and cohesion factors (Leech, 1994) may have affected word choice and frequently elicited function words in the LFP.

Similarly, in Fitzpatrick and Clenton's (2017) second study, to elicit participants' vocabulary knowledge in a non-discursive and more direct way, the BFP was used to make comparisons with Lex30 and to explore the extent to which the BFP and Lex30 scores are predictive of each other. The BFP test uses the same topic as the LFP, and students must write down as many single-word responses as possible. A new group of 80 (8 female, 72 male) L1 Japanese undergraduates were selected (TOEIC score 410-470 range). The results showed that the correlations between Lex30 and the BFP were not significant (r=.153 and r=.211).

Their third study compared Lex30 and G_Lex, which uses contextual priming and multiple prompts. Considering the lexical activation and other informing theories in their paper, Fitzpatrick and Clenton used G_Lex, which contains 24 sentences, with test-takers

required to write five different words for each sentence gap. Thus, G_Lex can elicit 120 words, the same maximum score as Lex30. The gaps to be filled in G_Lex are balanced among nouns, adjectives and verbs. G_Lex scores are calculated with the same frequency band as the previous two studies, and the G_Lex and Lex30 tasks showed significant correlations ($r=.645$, $p<.01$).

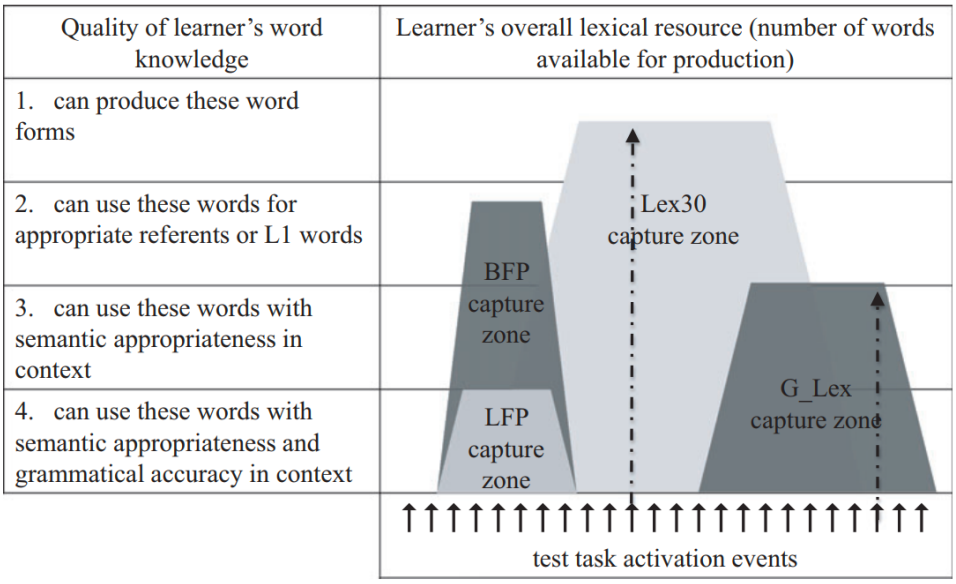
To interpret the productive vocabulary test results among the three experiments conducted in their paper, the authors discussed three matters imperative to effectively designing vocabulary tests. First, they addressed the advantages and frequency problems of the tests because all tests in their study utilised the frequency scoring system. The correlation scores among tests indicated that the tests do not tap into the same quality or quantity of word knowledge. Second, they wanted to reflect the conceptualisation of word knowledge according to the Vocabulary Knowledge Scale (VKS) (Paribakht & Wesche, 1993), a model applied for evaluating learners' word knowledge, which is essentially created for accessing word incidental vocabulary knowledge acquisition through reading. Increasing the four scales in VKS, five scales were adopted as the implicational scales as a quality dimension, and they also built in a quantitative dimension, including Lex30, the LFP, the BFP and G_Lex. They called it a Vocabulary Test Capture Model (see Figure 2.3). The model presented the similarities and differences between the four tasks used in their research. The model's vertical dimension is 'the quality of learner's word knowledge', the horizontal axis represents 'test activation events', and the learner's overall lexical knowledge is conceptualised above the horizon line. The LFP is a writing test requiring learners to use semantically and grammatically correct words. Lex30 requires students to write single words with no restrictions on grammar or syntax, which may tend to elicit more infrequent words.

Regarding the BFP, it requires word production with no semantic or grammatical regulations, but using the same topic as the LFP. It has almost the same horizon zone as the

BFP and a similar vertical dimension to Lex30. As for G_Lex, 24 sentences are divided into three semantic sections to elicit nouns, verbs and adjectives for every eight sentences, which requires students to produce words in the sentence contexts. G_Lex elicited less vocabulary than Lex30, but the sentence cues are apparently more stimulating than the LFP and the BFP, which have just a single essay title prompt. Conceptualising thoughts in and about these four tests can leave implications for productive vocabulary knowledge research. The model established in this study may influence future productive vocabulary task studies.

Figure 2.3

Vocabulary Test Capture: Lex30, the LFP, the BFP, and G_Lex



Note. Adapted from "Making sense of learner performance on tests of productive vocabulary knowledge," by T. Fitzpatrick and J. Clenton, 2017, *TESOL Quarterly*, *51*(4), p. 862. (https://doi.org/10.1002/tesq.356)

Critique

As mentioned at various points above, Lex30 is a task which leads learners to elicit more words in a general and broad way, whereas the LFP is an essay testing task measuring the vocabulary used with various restrictions, such as semantics, morphology, grammar, and collocation. In addition, learners have to consider how to deal with the topic and genre requirements during the writing process. With Lex30, there is no such burden. Even though Lex30 and the LFP are quite different, Fitzpatrick and Clenton's (2017) study used the same frequency-based approach to calculate both tests. I think this might be one reason for there being no significant correlations in the first experiment. More effective calculation methods need to be considered when accessing and assessing the vocabulary qualities of writing.

Fitzpatrick and Clenton's (2017) study also looked at Lex30 and G_Lex, which both use multiple prompts and cues to activate the lexical knowledge, and the number of responses is the same (120 English words). The differences between these two tasks are obvious. Lex30 uses single vocabulary cues, whereas G_Lex employs sentence prompts. There are no specific requirements for the responses to Lex30, whereas G_Lex requires only nouns, adjectives and verbs. Here an obvious problem arises when we put these responses into WebVP, which cannot deal with proper nouns, personal names, or names of places and countries and will thus automatically classify these words as being not on the list. Learners tend not to produce these words on G_Lex, but they will sometimes write these responses on Lex30, which may lead to some wasted responses. No direct comparison between either G_Lex and the LFP or G_Lex and the BFP was made in their paper. This may be a valuable direction for a more convincing Vocabulary Test Capture Model or future research.

In addition, the proficiency levels in all three experiments were the same. We still do not know if the Vocabulary Test Capture Model can also fit learners at various proficiency levels who need further validation. As this study mentioned, the four levels of learners' word knowledge on the vertical axis can be questioned; sometimes, language learners can use a

word correctly but cannot spell the word, especially since some words are acquired through incidental learning or some specific memorable context. For lower-level English learners, even a productive vocabulary task is the simplest (like Lex30), with no semantic or grammatical restrictions. However, students can often only produce highly frequent words or words which have no connection to the stimulus word or even show the wrong recognition of the cues. As Fitzpatrick (2006) discussed, word association tasks were originally used in psychology to express individual idiosyncrasies and some incomplete or inaccurate understanding of the stimulus or responses words can produce false cognates (Meara, 1984).

2.3 Review of Lexical Diversity Measure Studies

The current literature review section will now move on to evaluate four papers concerning lexical diversity measures. Treffers-Daller (2013) validated such more recently devised lexical diversity measures (D, HD-D, MAAS, and MTLD) and how these measures can predict participants' language proficiency. Treffers-Daller, Parslow, and Williams' (2018) paper investigated how lexical diversity measures can distinguish between participants of different proficiency levels, including by using both traditional and more recently developed lexical diversity measures. Vidal and Jarvis (2020) explored how lexical diversity measures can distinguish participants in a three-year-long English-medium instruction (EMI) education context. Kyle, Crossley, and Jarvis (2021) assessed the validity of lexical diversity measures through human raters' judgement.

2.3.1 Treffers-Daller, J. (2013): Measuring Lexical Diversity Among L2 Learners of French: An Exploration of the Validity of D, MTLD and HD-D as Measures of Language Ability.

Lexical knowledge is one of the main prerequisites for monolingual and bilingual children to achieve academic achievement (Daller, 1999; Dickinson & Tabors, 2001). Lexical diversity refers to the range of words used in a text, with a greater range showing a higher diversity (McCarthy & Jarvis, 2010). Treffers-Daller's (2013) paper provided valid information to measure lexical aspects of language ability using non-rating indices through speaking activities. Subsequently, Treffers-Daller proposed a crucial question in her paper to verify the validity of lexical diversity (LD) measurements by combining two more recently established LD measures, firstly with L1 French language learners for validation purposes, contributing to the overall construct of validity measurements.

The text length will quickly influence several LD measurements' ability to capture lexical knowledge of type (V) and token (N) ratios. The most well-known traditional LD measurement is the type-token ratio (TTR), which was pioneered by Johnson (1939; 1944) and Templin (1957). Other traditional LD measures include Mean Segmental TTR (Johnson, 1944), the index of Guiraud/Root_TTR (Guiraud, 1954), Log_TTR (Herdan, 1960), and the Maas index (Maas, 1972). D (vocd) measure, created by Malvern and Richards (1997), has been widely applied to assess LD in many languages. The more recently established measurements include the Measure of Textual Lexical Diversity (MTLD) by McCarthy (2005), HD-D (McCarthy & Jarvis, 2007), Moving Average TTR (MATTR; Covington & McFall, 2010), and MTLD-W (MTLD Wrap Around, Vidal & Jarvis, 2020). Both of these two more recently invented measures, MTLD and HD-D, needed further validation for other languages besides English; thus, Treffers-Daller's (2013) study was the first one validated for French.

Three groups of undergraduates took part in the study, in which group one comprised first-year undergraduates, group two was the final-year university students and group three was L1 French speakers, with the numbers of each group involving around twenty

participants. Their proficiency levels were group 1 (level 1) < group 2 (level 2) < group 3 (level 3). Participants were asked to retell two French comic stories and encouraged to prepare, and then their spoken production was recorded. Meanwhile, all participants completed the French C-test. For the lemmatisation process, all inflected forms of nouns, verbs and adjectives were lemmatised to their base form. In addition to articles, demonstratives, pronouns and question words, the study used masculine singular form throughout. CLAN calculated D values with both unlemmatised and lemmatised versions. With the FLO command's help in CLAN, we can convert unlemmatised and lemmatised versions into text format. MTLD, HD-D and MAAS scores were computed by Gramulator (McCarthy et al., 2012). The HD-D values results were all negative, with values closer to zero showing high diversity, and the values far below zero meant low diversity.

Treffers-Daller highlighted that the central issue in demonstrating MTLD and HD-D measures is concurrent validity, which is 'a criterion which we believe is also an indicator of the ability being tested' (Bachman, 1990, p. 248, as cited in Treffers-Daller, 2013, p. 82). The author chose the French C-test as the anchor test to predict to which degree the LD measures can assess participants' language proficiency. For validation construction, Treffers-Daller's paper mentioned four aspects: the effect of lemmatisation; predictive validity; internal validity; and convergent, discriminant/divergent, and incremental validity.

Treffers-Daller aimed to emphasise the importance of lemmatisation on the basis that French is highly inflected. The D value would be extremely high if the spoken production had not been lemmatised. Her paper used the base word as a word unit to analyse the data. The author contended that typological differences can be removed if the appropriate lemma principle is employed for different languages.

As presented in Table 2.9, the results of four LD measures are adjusted along with the lemmatisation process. The lemmatised data in her study can represent or explain the lexical

diversity scores and differentiate between the three groups of French learners. Table 2.10 uses the Eta-squared method to indicate the effect size of the lemmatised data and how strongly it can predict the particular group membership. The lemmatised data show a bigger effect size than the non-lemmatised data.

Table 2.5

Measures Calculated on Non-Lemmatised and Lemmatised Data (N = 64)

	Non-lemmatised M (SD)	Lemmatised (M, SD) M (SD)	t
MAAS	141.54 (15.53)	162.87 (16.85)	25.90**
D	41.95 (13.29)	26.98 (8.3)	19.82**
MTLD	40.27 (9.68)	30.64 (6.91)	14.19**
HD-D	–3.62 (2.19)	–6.07 (2.43)	20.06**

Note. **differences significant at $p < .001$. Adapted from "Measuring lexical diversity among L2 learners of French: an exploration of the validity of D, MTLD and HD-D as measures of language ability," by Treffers-Daller, J., 2013, In S. Jarvis and M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 79-103), p. 89. https://doi.org/10.1075/sibil.47.05ch3

Table 2.6

Effect Sizes (η2) of Measures Calculated for the Three Groups on Non-Lemmatised and Lemmatised Data (n = 64)

	Non-lemmatised data	Lemmatised data
HD-D	0.585	0.682
D	0.586	0.659
MAAS	0.362	0.429
MTLD	0.352	0.354

Note. Adapted from "Measuring lexical diversity among L2 learners of French: an exploration of the validity of D, MTLD and HD-D as measures of language ability," by Treffers-Daller, J., 2013, In S. Jarvis and M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 79-103), p. 90. https://doi.org/10.1075/sibil.47.05ch3

The C-test has been utilised as the anchor test to predict general language proficiency. According to the results in Table 2.11 and Table 2.12, significant correlations existed between the C-test and LD scores, which can be the predictive validity construct. Table 2.11 shows that the Pearson correlations between C-test and D and HD-D scores demonstrate closer relationships than those between the C-test and MTLD and MAAS scores, indicating that D and HD-D show better predictivity of learners' language ability than their MAAS and MTLD scores. Table 2.12 shows the results for sample sizes between 200 and 666 words, and the number of participants was 54.

Table 2.7

Correlations Between Measures of Lexical Diversity With the C-Test and Adjusted R^2 (N=64)

	MAAS[1]	D	MTLD	HD-D
Pearson r correlations with C-test (adjusted R^2)	–.556** (.298)	.763** (.575)	.571** (.326)	.791** (.620)

Note. [1]The correlation with MAAS is negative because low MAAS values indicate high diversity. Adapted from "Measuring lexical diversity among L2 learners of French: an exploration of the validity of D, MTLD and HD-D as measures of language ability," by Treffers-Daller, J., 2013, In S. Jarvis and M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 79–103), p. 91. https://doi.org/10.1075/sibil.47.05ch3

Table 2.8

Correlations Between Measures of Lexical Diversity and the C-Test, and Adjusted R^2 for Sample Sizes Between 200 and 666 (N=50)

	MAAS	D	MTLD	HD-D
Pearson r (adjusted R^2)	–.637** (.393)	.712** (.494)	.505** (.239)	.762* (.571)

Note. [1]The correlation with MAAS is negative because low MAAS values indicate high diversity. Adapted from "Measuring lexical diversity among L2 learners of French: an exploration of the validity of D, MTLD and HD-D as measures of language ability," by Treffers-Daller, J., 2013, In S. Jarvis and M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 79–103), p. 91. https://doi.org/10.1075/sibil.47.05ch3

The internal validity of sophisticated LD measures D, HD-D and MTLD has been studied by analysing their reliance on or resilience against text length. To validate if these three measures would work regardless of the text length, the study calculated different text length samples from the same texts, so if LD measures are indeed not dependent on the text length, the two segments' scores based on different text lengths would be the same as the scores for 300 words. However, the results in Table 2.13 and Table 2.14 show that the LD scores across different segments were different, meaning D, HD-D and MTLD values will differ for different text lengths.

Table 2.9

Mean and Standard Deviations for Lexical Diversity Scores Measured on Different Sample Sizes (N=30)

	D	HD-D	MTLD
100 words (mean of three segments)	30.19 (8.29)	–5.74 (1.76)	35.55 (7.88)
150 words (mean of two segments)	28.81 (8.18)	–5.70 (1.73)	34.60 (8.13)
300 words	31.39 (8.02)	–5.08 (1.65)	33.95 (7.76)

Note. Adapted from "Measuring lexical diversity among L2 learners of French: an exploration of the validity of D, MTLD and HD-D as measures of language ability," by Treffers-Daller, J., 2013, In S. Jarvis and M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 79–103), p. 92. https://doi.org/10.1075/sibil.47.05ch3

Table 2.10

Mean and Standard Deviations for Lexical Diversity Scores Measured on Different Segment Sizes (n=10)

	D M (SD)	HD-D M (SD)	MTLD M (SD)
140 words (mean of three segments)	32.20 (6.63)	–4.93 (1.47)	37.12 (6.11)
210 words (mean of two segments)	32.91 (6.77)	–4.77 (1.46)	37.41 (6.69)
420 words	35.29 (7.00)	–4.25 (1.42)	36.36 (6.81)

Note. Adapted from "Measuring lexical diversity among L2 learners of French: an exploration of the validity of D, MTLD and HD-D as measures of language ability," by Treffers-Daller, J., 2013, In S. Jarvis and M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 79-103), p. 94. https://doi.org/10.1075/sibil.47.05ch3 Copyright 2013 by the John Benjamins Publishing Company.

In addition, Treffers-Daller addressed incremental validity, convergent validity, and discriminant/divergent validity. Convergent validity is the concept that the theoretically similar construct measurements will be strongly correlated (Trochim, 2006). The perception of discriminant/divergent validity means that the theoretically different constructs could not be in high agreement (Campbell & Fiske, 1959). Incremental validity evaluates to what extent new measures can explain other measures (McCarthy & Jarvis, 2010).

Table 2.15 reveals the convergent validity results that D and HD-D scores showed strong correlations with each other, as MAAS and TTR scores showed that D and HD-D are similar in construct. Table 2.16 shows significant correlations between TTR and D and between HD-D and MTLD, indicating that LD measures are based on constrained text length (200–666). The limited text length increases correlation results between lexical diversity measures. Obvious changes (from no significant correlations to strong significant correlations) happen to the significant correlations between TTR scores and three sophisticated LD measures (D, HD-D, and MTLD).

Table 2.11

Correlations Between Measures of Lexical Diversity (n=64)

	D	HD-D	MTLD	MAAS	TTR
D	–	.93**	.77*	–.61**	.24
HD-D		–	.77**	–.62**	.22
MTLD			–	–.47**	.16
MAAS				–	–.85**
TTR					–

Note. Adapted from "Measuring lexical diversity among L2 learners of French: an exploration of the validity of D, MTLD and HD-D as measures of language ability," by Treffers-Daller, J., 2013, In S. Jarvis and M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 79-103), p. 95. https://doi.org/10.1075/sibil.47.05ch3

Table 2.12

Correlations Between Lexical Diversity Measures Calculated on Sample Sizes Between 200 and 666 (N=49)

	D	HD-D	MTLD	MAAS	TTR
D	–	.921**	.705**	–.763**	.575**
HD-D		–	.711**	–.771**	.551**
MTLD			–	–.503**	.369**
MAAS				–	–.915**
TTR					–

Note. Adapted from "Measuring lexical diversity among L2 learners of French: an exploration of the validity of D, MTLD and HD-D as measures of language ability," by Treffers-Daller, J., 2013, In S. Jarvis and M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 79-103), p. 96. https://doi.org/10.1075/sibil.47.05ch3

Table 2.17 shows the results of how accurately the respective LD scores can predict group membership. HD-D was the most informative one to measure the group membership, whereas MAAS and MTLD were weaker than HD-D and D, and the least successful one is TTR. When the text length changed to 200-666 words, MAAS became the most powerful index to predict the group membership, followed by HD-D and D, which showed that LD measures are easily influenced by text length, especially TTR scores.

Table 2.13

Group Membership as Predicted by Lexical Diversity Measures (Eta Squared)

	Eta squared (all samples, N = 64)	Eta Squared (samples from 200–666 words only) (N = 49)
HD-D	.682	.570
D	.659	.563
MAAS	.429	.593
MTLD	.354	.244
TTR	.253	.483

Note. Adapted from "Measuring lexical diversity among L2 learners of French: an exploration of the validity of D, MTLD and HD-D as measures of language ability," by Treffers-Daller, J., 2013, In S. Jarvis and M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 79-103), p. 98. https://doi.org/10.1075/sibil.47.05ch3

Critique

Treffers-Daller's paper investigated the functionality of the more recently invented LD measures in forecasting language learners' general language proficiency levels. The author mentioned four main aspects to validate these measures. Since French is a highly inflected language, the effect of lemmatisation should be considered. Incorporating the French C-test as an independent variable, her paper explored the predictive validity of LD

measures. Her findings showed that the most powerful measures are D and HD-D. To explore whether these more recently created measures depend on text length, she divided participants' spoken production into segments, and the results showed that all LD measures change with the sample length. The convergent, divergent and incremental validity results showed that the LD measures, based on the same construct, correlated strongly with each other regardless of text length. All these results provided a powerful reference for current LD studies and general language proficiency prediction, especially for French language learners.

Despite these strong points, some weaknesses remain. In the following critique, I address three concerns. These relate to (i) the lemmatisation only being mentioned in the results as a separate part (in the process of validity, the paper did not lemmatise the texts before calculating the data, which can influence the results of the calculations); (ii) some participants could be easily constrained by the speaking topics and genres, which would affect the length of their spoken production and vocabulary; and (iii) the author only included 10 participants whose production is equal to 420 words or more, and the sample size here is quite small scale for assessing the internal validity.

First, compared to Treffers-Daller et al.'s (2018) paper, Treffers-Daller (2013) used tokens as the lexical unit to verify the aspects of validity. Using the type as a word unit is important, as French is a high-inflection language. The obvious text length issue in the study is that different text lengths existed in the lemmatised and unlemmatised versions for participants' spoken production. If the lemmatised version is used for testifying the predictive, internal and divergent validities, the results may be different and we cannot compare the lemmatised and unlemmatised versions. A minor question here is that the author did not say how the data had been treated before calculating. As the test format is a story-telling task, we may assume that participants may prefer to use some words only used for speaking, non-existent words, unrecognisable words, and proper nouns. All these can

increase the LD scores because CLAN and Gramulator will treat these words as different types. The results here showed that sophisticated measures (D and HD-D) played a better role in predicting general language proficiencies, while Treffers-Daller et al. (2018) contradictorily revealed that simple measures (TTR and Types) work better than sophisticated measures.

Second, the speaking task used in Treffers-Daller (2013) was a story-telling speaking test, and students' output will be easily influenced by the topics. Some students are very talkative during the recording process, and it is voluntary to talk, while others may not be. The speaking topics and participants' characteristics may influence their production length and vocabulary. Different types of vocabulary would also influence LD scores.

Third, the sample size for participants who could produce longer text length was relatively small. To investigate the internal validity means evaluating to what extent the text length can influence sophisticated LD measures scores. Participants whose text length exceeds or exceeds 420 words were included, but only 10 met this requirement. Insufficient sample size for this specific group of participants may reduce the statistical power of the paper.

Treffers-Daller's paper is important and validated LD measures for L1 French participants. The findings in her paper showed that LD measures can predict language proficiency levels. Moreover, her paper emphasised that using lemmatisation and choosing consistent text length are two significant factors in generating accurate and valid LD scores. Despite these strengths in Treffers-Daller's (2013) paper, I have mentioned three potential concerns in her study: (i) the choice of word counting unit for validity; (ii) speaking topics may influence spoken production; and (iii) the number of participants who could produce longer text length was rather too small.

2.3.2 Treffers-Daller, J., Parslow, P., & Williams, S. (2018): Back to Basics: How Measures of Lexical Diversity Can Help Discriminate Between CEFR Levels.

Treffers-Daller, Parslow and Williams's (2018) research identified the importance of vocabulary knowledge in predicting writing scores and explored how researchers are keen to evaluate the vocabulary used in writing. Their paper pointed out that it is critical to distinguish learners' proficiency because language learners, particularly second language learners, require their writing to be assessed to determine different ways to improve their writing scores. Treffers-Daller et al.'s paper led the discussion by introducing various lexical diversity (LD) measures to assess the quality of essays written by second language learners whose levels were estimated to range from B1 to C2 of the CEFR (Common European Framework of Reference). The term lexical diversity usually connotes lexical richness (Read, 2000, p. 200), which involves four main aspects: *type-token ratio* or *lexical variation* (a variety of different words), *lexical sophistication* (the number of low-frequency words relating to the writing style and topic), *lexical density* (high percentage of lexical or content words compared to grammatical or function words) and the *number of errors* (few errors in the use of words). Lexical variation means the same as lexical diversity: namely, the range of expression and vocabulary knowledge necessary to avoid repetition. Malvern and Richards (2002) suggested that lexical diversity is the variety of active vocabulary deployed by a speaker or a writer. In addition, Jarvis (2013a) proposed a perception-based phenomenon with six measurable properties to define the lexical diversity construct.

Treffers-Daller et al.'s (2018) study comprised five sections: (i) section one is the introduction; (ii) section two presents the LD construction and measurements; (iii) section three outlines the different definitions of words relating to LD; (iv) section four presents their methodology, and the results; and (v) section five concludes their study and provides implications for future research.

Their first section introduced their research and aimed to look at how different basic units of measurement, namely, the word, the lemma or the word family, affect the LD measures scores and their power to predict CEFR levels. Then, the construction and measurements of LD must be clarified before using it to evaluate writings. Treffers-Daller et al. outlined LD as being not just about the range of words but also about how these words are deployed in texts (Laufer & Nation, 1995; Durán et al., 2004). Treffers-Daller et al.'s paper refers to Jarvis's (2013a) research that discussed six dimensions or properties of the LD construct: variability (inherent property of redundancy); volume (vocabulary size); evenness (balance); rarity (less common/frequent words); dispersion (spatial distribution); and disparity (degree of differentiation). Treffers-Daller et al.'s (2018) paper highlighted that they would only focus on one aspect of LD: variability. Treffers-Daller et al.'s (2018) paper used both traditional LD measures and several more recently devised measures. Their paper categorised that traditional measures include TTR (Type-Token Ratio) (Johnson, 1944; Templin, 1957), whereas the more recently developed LD measures mainly comprise D (Malvern et al., 2004), MTLD (McCarthy, 2005) and HD-D (McCarthy & Jarvis, 2007).

One important factor noted by Treffers-Daller et al. was that these measures, whether simple or complex, are influenced by text length. The longer the texts are, the lower the LD scores with TTR and MTLD, whereas the D and HD-D values will increase along with the growth of text length. Therefore, the authors keep all the writing samples to 200 words to control for text length.

Another crucial issue relates to Treffers-Daller et al.'s discussion of 'What is a type?' According to this paper, we can conclude mainly three ways to treat a word (type). Numerous scholars consider the different tokens of inflected forms as the same type. Durán et al. (2004) classed fused forms (such as *fell-fall*) as distinct types. On the contrary, Yu (2010) and VocabProfile (www.lextutor.ca/vp/comp/) considered all the inflected forms as distinct types.

Laufer and Nation (1995) (reviewed in section 2.2.1) used all inflected forms and derived forms up to the level three proposed in Bauer and Nation (1993): namely, word family classifications. From the psycholinguistic perspective, Treffers-Daller et al. (2018) suggested that L2 learners have difficulties producing derived forms of a word based on the root form, especially for morphologically complex words, which indicates it might not be the best idea to consider the derived forms as one type.

An innovative aspect of Treffers-Daller et al.'s research related to their discussion of how different definitions of *word* influence LD values. Their paper proposed three different lemmatisation standards: non-lemmatised words (lemma 0); all the inflected forms, including verbs, nouns, and adjectives counting as same tokens (lemma 1); and word families including all the inflected forms and derivational affixes up to level 3 (lemma 2). To explore these three different approaches to lemmatise their study's data, they presented two fundamental research questions: First, what is the effect of different types of lemmatisation on the LD scores; second, how do different lemmatisation principles affect the ability of the LD measures to discriminate between CEFR levels?

Regarding their study section, the participants in Treffers-Daller et al.'s paper were 179 adults from 42 different countries, and their writing samples were from the Pearson Test of English Academic (PTE Academic). In addition, the PTE Academic offers writing scores, CEFR levels, vocabulary scores and overall scores for all participants.

Table 2.14

Students' Level of Competence According to the CEFR

CEFR level	B1	B2	C1	C2
N	50	50	50	29

Note. Adapted from "Back to basics: How measures of lexical diversity can help discriminate between CEFR levels," by Treffers-Daller, J., Parslow, P., and Williams, S., 2018, *Applied Linguistics*, *39*(3), p. 310. Copyright 2018 by the Oxford University Press.

Treffers-Daller et al.'s paper adopted six different LD measures across these three different types of lemmatisation. The six LD measures are Types, TTR, Guiraud, D (vocd), HD-D and MTLD, in which CLAN (MacWhinney, 2000) plays an important role in different lemmatisation versions and calculates D scores. Usually, they suggested that students who have higher CEFR levels have a greater variety of words. Their results showed that both basic measures of LD (Types, TTR and Guiraud) and sophisticated measures (D, HD-D and MTLD) can distinguish participants across four CEFR levels (B1, B2, C1 and C2).

Their results, shown in Table 2.19 and Table 2.20, highlight the differences between LD scores based on different lemmatisation standards; using lemmatisation standards (lemma 1 and lemma 2) would lower LD scores. Basic LD measures proved better able to predict the CEFR levels more consistently than the sophisticated measures, and the correlations among basic LD scores demonstrate closer relationships than among the sophisticated measures. They concluded that the lemma 1 principle can be more useful in discriminating students' levels than lemma 2. Treffers-Daller et al.'s (2018) paper further revealed that important knowledge will be missed if we remove the derived vocabulary forms, especially for L2 English language learners. Meanwhile, the regression analyses in their paper also indicated

that among all LD scores. *Type* shows better predictivity of overall scores, writing scores and vocabulary scores than the rest. Specifically, *type* can explain 22 percent of the variance in overall scores, 20 percent in writing scores, and 21 percent in vocabulary scores.

Table 2.15

Basic Measures of Lexical Diversity Across Different Levels of the CEFR

Measures	B1	B2	C1	C2	Overall means and SD	Eta squared
Types 0	101.52	109.48	111.66	114.76	108.72 (9.98)	.225
Types 1	96.32	104.14	106.32	109.48	103.43 (9.82)	.229
Types 2	96.24	103.92	106.06	109.07	103.21 (9.87)	.221
TTR 0	0.56	0.61	0.61	0.63	0.60 (0.06)	.229
TTR 1	0.52	0.57	0.58	0.60	0.60 (0.06)	.248
TTR 2	0.52	0.57	0.58	0.59	0.56 (0.06)	.234
Guiraud 0	7.51	8.14	8.27	8.48	8.05 (0.74)	.232
Guiraud 1	7.09	7.71	7.86	8.08	8.03 (0.74)	.242
Guiraud 2	7.08	7.69	7.84	8.04	7.50 (0.73)	.230

Note. Adapted from "Back to basics: How measures of lexical diversity can help discriminate between CEFR levels," by Treffers-Daller, J., Parslow, P., and Williams, S., 2018, *Applied Linguistics*, *39*(3), p. 315. Copyright 2018 by the Oxford University Press.

Table 2.16

Sophisticated Measures of Lexical Diversity Across Different Levels of the CEFR

Measures	B1	B2	C1	C2	Overall means and SD	Eta squared
D (VOCD) 0	72.4	85.71	86.61	89.54	82.86 (21.29)	.092
D (VOCD) 1	61.88	71.65	73.83	76.61	70.33 (17.28)	.098
D (VOCD) 2	62.2	71.58	74.48	75.67	70.15 (17.25)	.085
HD-D 0	34.47	35.37	35.51	35.64	35.21 (1.40)	.109
HD-D 1	33.55	34.29	34.55	34.75	34.23 (1.39)	.100
HD-D 2	33.61	34.36	34.36	34.86	34.30 (1.40)	.104
MTLD 0	70.14	84.55	88.47	93.85	83.12 (22.96)	.134
MTLD 1	58.7	68.52	72.81	77.11	68.37 (17.06)	.140
MTLD 2	59.68	70.01	73.69	78.92	69.60 (17.82)	.145

Note. 0=no lemmatisation, 1=first lemmatisation principle, 2=second lemmatisation principle. Adapted from "Back to basics: How measures of lexical diversity can help discriminate between CEFR levels," by Treffers-Daller, J., Parslow, P., and Williams, S., 2018, *Applied Linguistics*, *39*(3), p. 316. Copyright 2018 by the Oxford University Press.

Critique

The findings in Treffers-Daller et al.'s (2018) paper reveal that both basic measures (Types, TTR, and Guiraud) and sophisticated measures (D, HD-D, and MTLD) can predict CEFR levels (B1 to C2) and different lemmatisation standards affect the LD scores. Treffers-Daller et al. used three lemmatisation standards: no lemmatisation, lemma-based lemmatisation, and word family-based lemmatisation. The lemma-based standard can more easily distinguish the CEFR levels than the other two. The Eta squared values indicate basic LD measures show better predictions of CEFR levels than sophisticated LD measures. Correlations between the basic LD measures and overall scores, writing, and vocabulary scores demonstrated closer relationships than the sophisticated measures. Despite these strengths, we should discuss some problems in Treffers-Daller et al.'s (2018) study. In the

following critique, I address three concerns: (i) some measures of different lemmatisation cannot distinguish between B2 and C1 level; (ii) the participants in their paper are of mixed language background, but for Indo-European language learners and non-Indo-European language learners, different lemmatisation principles should be used; (iii) the calculation procedures use CLAN to lemmatise texts.

First, the concern in Treffers-Daller et al.'s paper is that some measures cannot discriminate the CEFR scores, as shown in Table 2.19 and Table 2.20. At B2 and C1 levels, TTR scores are the same, which cannot distinguish students' proficiency. Regarding the sophisticated measures in Table 2.20, the HD-D scores cannot distinguish the levels for participants belonging to B2, C1 and C2. As shown in Table 2.21, based on the first lemmatisation principle, the levels among B1, B2, C1 and C2 have significant differences, especially between the lowest level B1 and higher levels (C1 and C2). The LD scores for the B2 level show no significant difference between C1 and C2. Table 2.21 only referred to their first lemmatisation scores. It did not mention their second lemmatisation LD scores, but I think is better to address the significant difference between LD scores and the CEFR levels with the lemma 2 standard.

Table 2.17

ANOVA and Tukey Post Hoc Test Results for Lexical Diversity Measures (First Lemmatisation Principle) Across Different Levels of the CEFR

	F	p	B1–B2	B1–C1	B1–C2	B2–C1	B2–C2	C1–C2
Types	17.034	<.0001	*	*	*	NS	NS	NS
TTR	18.923	<.0001	*	*	*	NS	NS	NS
Guiraud	18.27	<.0001	*	*	*	NS	NS	NS
D (VOCD)	6.198	.0005	NS	NS	*	NS	NS	NS
HD-D	6.388	.0004	NS	NS	*	NS	NS	NS
MTLD	9.757	<.0001	NS	*	*	NS	NS	NS

Note. * means significant difference between CEFR levels. For post hoc comparisons, alpha was set at .0014. Adapted from "Back to basics: How measures of lexical diversity can help discriminate between CEFR levels," by Treffers-Daller, J., Parslow, P., and Williams, S., 2018, *Applied Linguistics*, *39*(3), p. 317. Copyright 2018 by the Oxford University Press.

Second, an additional concern is using the same lemmatisation standard with participants of different language backgrounds. The 176 participants in their paper were from 47 countries, and it is hard to know the role of English in their lives. We cannot know whether these learners are all L2 or L1 English language learners. Besides, considering their participants in slightly more detail with regard to specific L1 backgrounds may need different considerations about their lemma standards. As with L1 participants, especially with participants familiar with a similar orthographic system, these participants may have a distinct advantage in learning morphological knowledge over L2 participants. Regarding L2 participants, it is obvious that the derived vocabulary forms are part of their language ability. It is a significant step to use appropriate lemmatisation standards, such as lemma/flemma, to distinguish their CEFR levels.

Third, another concern is that Treffers-Daller et al.'s (2018) research used CLAN to lemmatise the writings. However, CLAN only computes D(vocd) scores, and the code in CLAN cannot be recognised by other software. The authors must build both lemma one and lemma two writing samples to get other LD scores. This process has not been explained. If all the writings for three different lemmatisation standards are created manually, it is a time-consuming process. Another issue with CLAN is that it cannot fully lemmatise all the words at the morphosyntactic tier (the analysis interface presented in CLAN for the lemmatised words). Take the word 'being' as an example: some uses of 'being' cannot be lemmatised as 'be'. Thus, to some extent, CLAN is an imperfect tool for the lemmatisation process.

Treffers-Daller et al.'s (2018) study is important because it measures how LD scores can distinguish between CEFR levels across different lemmatisation standards. The findings of their paper showed not only that LD scores can differentiate several CEFR levels but also that there are significant correlations with vocabulary scores, writing scores, and overall scores. Treffers-Daller et al.'s (2018) study has implications for future research, using LD measures to distinguish writing and CEFR levels. In addition, their study shows implications for future studies into collocations or different word units because collocations are another important factor influencing learners' language proficiency and writing scores. All the lexical diversity measures only estimate single words and do not combine other words that appeared simultaneously, which I think needs further development regarding the evaluation of LD scores and its related words, not only single words, to produce more objectivity and accuracy in judging learners' lexical proficiency in their writing. Despite these strengths in Treffers-Daller et al.'s (2018) paper, I have also mentioned three potential weaknesses in their study: (1) some LD measures cannot distinguish between CEFR levels, and it only shows lemma1 results in differentiating CEFR levels; (2) lemmatisation standards should consider

participants with different language backgrounds; and (3) the lemmatisation process with CLAN can be inaccurate.

2.3.3 Vidal, K., & Jarvis, S. (2020): Effects of English-Medium Instruction on Spanish Students' Proficiency and Lexical Diversity in English.

English-medium instruction (EMI), a means to increase the world ranking for universities and sharpen students' competitiveness in language skills relating to their future careers, has been adopted by many modern universities. Vidal and Jarvis's (2020) paper demonstrated that an EMI trend (Earls, 2016, p. 2) existed among modern higher education institutions, as in Spain. However, EMI is a controversial issue, as stated by Macaro (2017, p. 2), for its potential to eliminate cultural diversity (Wilkinson, 2013) or obstruct the depth of knowledge learning (Airey, 2015).

In Vidal and Jarvis's paper, they aimed to examine the consequences of a three-year-long education under EMI and to explore the relationships among the English proficiency of learners, writing quality, and lexical diversity (LD), which were respectively and separately measured by the Oxford Placement Test (OPT), the CEFR writing band descriptors (Council of Europe, 2001), and three LD measurements. In their study, they tested 195 undergraduates at two different proficiency levels in a Spanish university. Their study's results implied an increase in L2 learners' language proficiency levels and a slight improvement in writing quality but no significant improvement in their LD scores. Vidal and Jarvis also pointed out that studies about language learning through EMI at the higher education level were very rare, and the issue of whether EMI lessons can improve the students' language abilities is not transparent.

Due to the rarity of EMI studies concentrating on vocabulary usage, Vidal and Jarvis reviewed a similar concept, the academic investigations of Content and Language Integrated

Learning (CLIL), to reveal vocabulary acquisition studies of CLIL. Some researchers found CLIL students produced larger TTR (type-token ratio) values (Llach & Catalán, 2007) and appeared to have larger vocabulary sizes, both receptive and productive, and more knowledge of less frequent words and greater vocabulary accuracy (Dalton-Puffer, 2011), but two studies (Catalán & Agustín Llach, 2017; Roquet & Pérez-Vidal, 2017) found no differences between the CLIL group and non-CLIL group.

Vidal and Jarvis also wanted to address whether students' LD knowledge could develop along with their three-year-long EMI lessons. Studies (Jarvis, 2017; Treffers-Daller, 2013) reported significant relations between LD and learners' language proficiency. Jarvis (2017) explained that current LD measures evaluated lexical repetition, an *etic* perspective of language, which lacked construct validity. They also observed that Zipf (1935) implied that capable language speakers shared a similar perception of language by considering LD as a matter of perception. Thus, Jarvis (2017) adopted many human raters, which proved Zipf's statements. They demonstrated that LD had more connections with redundancy (a psychological phenomenon) than repetition, and the construct of LD measurements was multidimensional and was still under development (Jarvis, 2013a, 2013b, 2017). For the purpose of evaluating their study, Vidal and Jarvis adopted three LD measures, namely the Measure of Textual Lexical Diversity (MTLD) (McCarthy, 2005), Moving Average Type-Token Ratio (MATTR) (Covington & McFall, 2010), and MTLD-W.

Two groups of students joined their studies, 49 students in their first academic year and 59 students in their third academic year, and all of them were undergraduates whose major was English. An argumentative essay was assigned to all participants with a general topic based on the TOEFL test to elicit more ideas and the language ability of their writing. OPT was used as a standard level test to assess students' proficiency levels for first-year and

third-year students. However, the OPT results showed no significant differences in proficiency levels between the two groups.

Vidal and Jarvis annotated the POS (parts of speech) regarding the data processing process. They lemmatised for all essays by using Tree Tagger (Schmid, 1994, 1995), a free and widely used tool reaching 96.36% accuracy for English (Schmid, 1994), during which all the words of Latin origin or the mixed usage of Latin words with English in their essays were not included in the lemmatisation process and were annotated with a specified prefix for additional analysis. Three LD measures were adopted: MTLD, MATTR, and MTLD Wrap Around (MTLD-W). MTLD was validated to correlate with learners' language abilities moderately (Treffers-Daller, 2013) and their writing levels (McCarthy & Jarvis, 2010). The CEFR Writing Scale (Council of Europe, 2001) was used as a reliable measure to assess learners' writing skills (Huhta et al., 2014; Zheng et al., 2016), and each essay was graded with a number from one to six, with the higher significant number meaning a higher writing level. One hundred and eight writings (49 for first-year students and 59 for third-year students) were judged by 39 human raters who were graduate students from two universities and were given the rubric and sample essays as a training process. Professional raters rated all the sample essays, and their inter-rater agreement values reached 0.852 and 0.849, respectively.

Their results showed that no significant differences existed among MTLD, MTLD-W and MATTR values between first-year EFL learners and third-year EFL students, which answered Vidal and Jarvis's first research question (whether the third-year students showed a higher LD than first-year students after three-year-long studies under EMI instruction), showing that there was no improvement in their LD scores after three-year-long EMI lessons. Regarding question two (whether students' English proficiency had significant correlations with their LD scores), students' proficiencies estimated by OPT correlated with their LD

scores. The strongest correlation was with MTLD-W (r=.36), followed by MTLD (r=.35) and MATTR (r=.32). Meanwhile, their mean OPT scores were M (1st-year students)=73.84; M (3rd-year students)=79.11. After converting these scores into their corresponding CEFR levels, students' levels improved from B2 to C1 level, implying that their language ability had been improved, which answered question three (whether third-year level students had higher language proficiency than first-year students). Regarding question four (comparing students' writing proficiency between first-year and third-year students), students' writing levels were measured by CEFR writing scales (see Table 2.22); the results showed significant differences between 1st-year students and 3rd-year students on their writing qualities. The third-year students' writing quality had improved, as judged by CEFR raters. However, it should be noted that their writing levels remained at B1 level, which meant that although their writing level had improved, it still fell behind students' general language proficiency as measured by OPT. As for question five (whether participants' essay qualities correlated with LD), results showed weak significant correlations between students' writing levels and their LD scores. For instance, the correlation between writing level and MTLD-W was r=.33, and the correlation between writing level and MATTR was r=.31. The correlation between writing level and MTLD was r=.30. As for question six (whether students' language proficiency correlated with their essay quality), results showed moderately significant correlations between general language proficiency (OPT) and essay quality (writing scores based on CEFR writing scale) (r=.58, p<.001).

Table 2.18

1st-Year and 3rd-Year Essay Quality (n=109)

Year	n	Mean	Standard deviation	Minimum	Maximum	α
1	49	3.55	.61	2.58	5.11	0.852
3	59	3.80	.55	2.64	5.14	

Note. Adapted from "Effects of English-medium instruction on Spanish students' proficiency and lexical diversity in English," by K. Vidal, and S. Jarvis, 2020, *Language Teaching Research*, *24*(5), p. 12. Copyright 2020 by the SAGE Publications.

Vidal and Jarvis discussed and concluded the possible reasons why students' LD knowledge had not been improved (see Table 2.23). First, they believed students were more exposed to the vocabulary relating to their field of study and academic words. Second, they mentioned that much receptive vocabulary under EMI education could not fulfil practical productive usage in participants' writings. Third, they emphasised EMI courses in writing were much more for content and organisation of essays than for selecting a greater variety of words to compose essays.

Table 2.19

Mean Lexical Diversity Scores (With Standard Deviations in Parentheses)

Year	MTLD	MTLD-W	MATTR
1	65.61 (13.71)	63.71 (12.86)	38.44 (1.72)
3	63.27 (12.35)	63.96 (13.75)	38.42 (1.49)

Notes. MTLD = Measure of Textual Lexical Diversity; MATTR = Moving Average Type-Token Ratio. Adapted from "Effects of English-medium instruction on Spanish students' proficiency and lexical diversity in English," by K. Vidal, and S. Jarvis, 2020, *Language Teaching Research*, *24*(5), p. 12. Copyright 2020 by the SAGE Publications.

Vidal and Jarvis also discussed why no significant correlations existed between students' general language ability and LD measurements. They explained that even though the third-year students' CEFR proficiency assessed by OPT increased from B1 to C1 level, their LD scores did not appear significantly different from their first-year counterparts. Thus, they argued that proficiency improvement was not significant enough for an increase in LD, which meant a high proportion increase in LD might but not definitely improve language ability. Regarding the relationships between writing quality and LD, three LD measures showed weak significant correlations. In their paper, they also mentioned that Yu's (2010) paper also got weak correlations with the D measure (r=.29), which was similar to their results for three LD measurements: MTLD, MTLD-W and MATTR (r=.30-.33). Vidal and Jarvis concluded that students' writing proficiency lags behind their other language skills. It was worth considering L2 communicative development and language skills, such as vocabulary skills (Zheng et al., 2016), which were not clearly explained on the CEFR writing scale. They showed that improving the quality of input in students' lectures (Airey, 2015) and

overloading content with English as a medium of instruction may cause their working memory to become exhausted by dealing with their English language learning. Thus, their practical studies provided evidence for language policy and objective makers at the university level.

Critique

Vidal and Jarvis's paper mainly investigated language skills development at the higher education level in Spain in the English-medium instruction (EMI) context. Their research focused on L2 learners' general language proficiency, writing qualities, and lexical diversity. Their results showed that students' L2 proficiency significantly improved, and their writing quality improved, but no significant differences existed among their LD scores. Their results and findings suggested reassessing and reconsidering the enrolment standard and quality of EMI lecturers at the higher education level. Despite the strengths of their article, there are still four potential concerns.

The first concern relates to the Oxford Placement Test (OPT) scores. As described in Vidal and Jarvis's study, OPT comprises two parts: grammar and listening. However, their study only used grammar testing to assess students' language proficiency. They used students' grammar scores to reflect their general language ability standards, even though their overall language ability cannot entirely be reflected only by their grammar scores. Thus, finding a rational criteria test to assess students' language proficiency is reasonable and necessary for future data analysis.

The second concern relates to human rating descriptors. In Vidal and Jarvis's study, they used 39 human raters to give scores to all essays based on the CEFR writing scale, which does not mention lexical usage, meaning the raters neglect the students' lexical

features in their essays while rating them, unlike in the TOEFL and IELTS writing band descriptors, which address the lexical aspect of writing specifically.

The third concern is that the test-takers in their study comprised two different numbers and groups of students; specifically, there were 49 first-year and 59 third-year students. As they explained in their paper, Vidal and Jarvis's study was longitudinal research. However, they adopted two different groups of students, and the total number of students was also different. Students could have had different English levels at the time of their university enrolment, and the English language lectures they took were also different. Thus, their study cannot be considered a rigorous longitudinal study. Future investigations about LD measures, CEFR writing scales, and L2 proficiency assessment in the EMI context should be further executed and evaluated either cross-sectionally or longitudinally.

The fourth concern is the LD measures; Vidal and Jarvis only use three measures in their paper, namely MTLD, MTLD-W, and MATTR. In Treffers-Daller's (2013) study, she found that lexical diversity measures, such as HD-D and D, play a vital role in predicting students' general language ability by reaching 62% predictive ability of students' L2 proficiency scores. Supposing their research had adopted a wider variety of LD measures, the results, such as for HD-D and D, may be different.

Vidal and Jarvis's (2020) paper offered insights into participants' proficiency in the EMI education context. They proposed a more recently established LD measure (MTLD_W) and used three (MTLD, MTLD_W, and MATTR) to predict participants' improvement after a three-year-long EMI program. Despite these strengths in their paper, I have highlighted some potential concerns in my critique: (i) the OPT scores for general language proficiency cannot represent participants' overall language proficiency; (ii) the rating descriptors neglect lexical features during raters' judgement process; (iii) different groups of participants of

different levels cannot be considered a strictly rigorous longitudinal study; and (iv) their study may be limited by only using three LD measures.

2.3.4 Kyle, K., Crossley, S. A., & Jarvis, S. (2021): Assessing the Validity of Lexical Diversity Indices Using Direct Judgements.

Lexical diversity measures have long been applied to assess vocabulary size (Jarvis, 2002, 2013b) and proficiency in speaking, writing or language curriculum level (e.g., Engber, 1995; Jarvis, 2002; Treffers-Daller et al., 2018). Text length issues are a typical problem among lexical diversity indices, and researchers have dedicated themselves to the improvement of the measurements to maintain a stable value across different text lengths (e.g., Covington & McFall, 2010; Malvern & Richards, 1997; McCarthy & Jarvis, 2007, 2010). Very few studies have focused on the extent to which human ratings of lexical diversity can help validate lexical diversity indices. In Kyle, Crossley and Jarvis's (2021) paper, they mainly investigated three dimensions of lexical diversity: abundance, variety, and volume. Kyle et al.'s study also used human rating scores in learners' argumentative essays corpus. Their results revealed that abundance could predict the scores of lexical diversity by human ratings most, and abundance and variety could predict around 74% of lexical diversity ratings.

In their paper, Kyle et al. indicated that the text length problem has been widely realised, but much less is known about which lexical diversity measures can best construct or measure the lexical diversity. In addition, no studies had investigated lexical diversity scores in argumentative essays though lexical diversity measures have been used in assessing writing tasks for many years. Moreover, their paper explored the participants of both L1 and L2 language backgrounds. The authors also mentioned that conventionally calculating lexical

diversity requires computer knowledge, which may also block the construct process development in this area.

In addition, Kyle et al. corrected the misunderstanding that lexical diversity measures should be treated as multidimensional events, not unidimensional. According to Jarvis (2013a, 2013b, 2017), seven characteristics of lexical diversity are introduced, such as *volume* (tokens), *abundance* (types), *variety* (ratio of unique words), *evenness* (the degree of equal repetition many types), *disparity* (semantic relationship of words), *specialness* (the appearance of particular words to increase diversity), and *dispersion* (the distance between the recurrence of the same words). Kyle et al.'s paper only focused on the first three dimensions: volume, abundance, and variety, because they stated that the scope of the paper and the measures of these dimensions still needed further development.

Their paper resolved two main issues of lexical diversity measures. First, they indicated that many lexical diversity indices are expected to demonstrate both lexical variety and vocabulary size. The most widely known LD measure is TTR (type-token ratio) (Johnson, 1944), the total types divided by the total tokens in a text. Their paper mentioned that the issue with TTR is that the longer the text, the lower the lexical diversity value. Kyle et al. (2021) emphasised two reasons causing the TTR problem. One is that more proficient learners will be more fluent (able to complete longer tests in a short time); thus, adopting a larger vocabulary and TTR index cannot test the learners' real vocabulary knowledge. Some indexes, like Guiraud, have positive relations with text length, and it will cause extremely high diversity scores (Koizumi & In'nami, 2013). Their paper also stated that text length positively influences human ratings, and human raters give higher scores to longer texts. The second concern is the exactness of lexical diversity measurement; theoretically, lexical diversity measures should only indicate diversity regardless of other textual or user features. However, the textual features influence the text length. If the lexical diversity measures are

inherently influenced by text length, we cannot say if they accurately reflect the lexical diversity or other textual characteristics.

To address the abovementioned issues involving lexical diversity measures, Kyle et al. used multiple lexical diversity measures (the first three dimensions of lexical diversity) and human rating judgement in evaluating each essay.

In Kyle et al.'s paper, they examined two corpora with the same writing genre (argumentative essays), including two groups of participants: L1 undergraduates with 315 essays and L2 TOEFL test-takers with 300 writings. Two trained lexical diversity raters graded all the essays, and their inter-rater agreement value was .748 after adding an extra adjudication opportunity.

They used the Tool for the Automatic Analysis of Lexical Diversity (TAALED), which is open source and free for lexical diversity indices except for the D (vocd) measure. They used TAALED to investigate the three aspects of lexical diversity pointed out by Jarvis (2013a, 2017): volume, abundance, and variety. Volume means the total number of words (tokens) in the text. Abundance expresses different types (of lemma) in writing. For variety, four measures which are least influenced by text length have been used: HD-D (McCarthy & Jarvis, 2007), MATTR (Covington & McFall, 2010), MTLD (original) (McCarthy & Jarvis, 2010) and MTLD-W (McCarthy & Jarvis, 2010).

Regarding the data computing process, Kyle et al. calculated bivariate Pearson correlations for their first research question to explore the relationships between lexical diversity indices and human rating scores. They used a linear model for their second research question to predict human rating scores through lexical diversity measures.

Their results showed moderate to strong correlations between lexical diversity indices (variety, volume and abundance) and human ratings of lexical diversity. Table 2.24 shows the correlations between lexical diversity indices and the human judgements of lexical diversity

scores. They concluded that abundance had the strongest significant correlations with human rating scores, both with the combined corpus (r=.847) and separate L1 or L2 corpora, and volume and other indices came in behind, which suggested that human rating scores accepted text length as a factor.

Table 2.20

Correlations Between Lexical Diversity Indices and Human Judgements of Lexical Diversity

	Combined		L1		L2	
Index	*r*	*p*	*r*	*p*	*r*	*p*
Volume	0.687	<.001	0.683	<.001	0.695	<.001
Abundance	0.847	<.001	0.815	<.001	0.890	<.001
MATTR	0.492	<.001	0.402	<.001	0.566	<.001
HD-D	0.602	<.001	0.522	<.001	0.666	<.001
MTLD	0.505	<. 001	0.438	<.001	0.566	<.001
MTLD-W	0.524	<.001	0.433	<.001	0.612	<.001

Note. Adapted from "Assessing the validity of lexical diversity indices using direct judgements," by K. Kyle, S. A. Crossley, and S. Jarvis, 2021, *Language Assessment Quarterly*, *18*(2), p. 9. Copyright 2021 Taylor & Francis.

As for the second research question, Kyle et al. used linear models among human rating lexical diversity scores, abundance, speaker status, and lexical diversity indices. Table 2.25 summarises all regression models indicating that these four models explained approximately 74% of human rating lexical diversity scores. The lexical diversity measurement of abundance had the most power to predict each model, and the various

indices could also explain the model. The authors concluded that objective lexical diversity measurements formulated descriptive models of human ratings of lexical diversity.

Table 2.21

Summary of Regression Models

Model	$R^2_{adjusted}$	Relative importance of abundance	Relative importance of variety index	Relative importance of other features
Abundance + MATTR	0.736	0.588	0.121	0.029
Abundance + HD-D	0.737	0.532	0.181	0.026
Abundance + MTLD (original)	0.735	0.580	0.127	0.030
Abundance + MTLD-W	0.735	0.571	0.136	0.030

Note. Adapted from "Assessing the validity of lexical diversity indices using direct judgements," by K. Kyle, S. A. Crossley, and S. Jarvis, 2021, *Language Assessment Quarterly*, *18*(2), p. 13. Copyright 2021 Taylor & Francis.

Critique

Kyle et al.'s study explored the relations between human rating scores of lexical diversity and three dimensions of lexical diversity: volume, abundance and variety. Jarvis

(2013a, 2013b, 2017) proposed the important role of human judgement in argumentative essays with both L1 and L2 English language learners. Kyle et al.'s research results showed abundance (number of types) strongly correlated with holistic human rating scores in two corpora. The established models could explain approximately 74% variance in human judgements of lexical diversity scores. All these findings are beneficial for comprehending lexical diversity indices and human judgement of lexical diversity when confronting new ideas. Despite these strong points, their paper includes three potential limitations.

First, the problem relates to text length. Kyle et al.'s research included all running words of participants' essays. Two corpora in their research include L1 speakers' SAT writings and L2 speakers' TOEFL writings. However, lexical diversity scores were claimed to be independent of text length, as investigated by Treffers-Daller's (2013) paper, but in fact, they would be influenced by text length. The abundance (the number of types), the strongest predictor, will increase along with text length, especially when learners have more ideas or opinions about a certain writing topic. Kyle et al. argued that human raters will also give higher scores while dealing with longer essays. One reason might be that participants who can produce longer texts represent those who possess higher language proficiency than participants who cannot. Human judgements include many subjective factors, and we cannot find evidence that they all prefer long texts. In addition, this preference may not grow or become manifest a commensurate at the same rate as the increasing number of types.

Second, another problem relates to the lemmatisation process. The polysemous words in their writing corpora should be distinguished if their paper uses accurate lemma standards. In an earlier study, Kyle (2019) had stated that many papers purporting to use lemmas are actually flemmatising the texts. However, Kyle et al. (2021) did not explain this problem. Many investigations (e.g., McLean, 2018; Nation, 2021) have investigated whether we need

to use lemma, flemma or word families to best gauge particular learners' language proficiency.

Third, as the research results showed in Kyle et al.'s paper, abundance is the prime factor in predicting lexical diversity in argumentative writings. At the same time, there are still numerous unknowns about other written or oral tasks. In light of the construction development question pointed out by Jarvis (2013a, 2013b, 2017), Kyle et al.'s research only investigated three components of lexical diversity. More measures will be validated in due course, and then we can choose appropriate lexical diversity measurements based on different genres or research questions.

Kyle et al.'s (2021) paper is important because it evaluated how lexical diversity measures relate to the human judgment of lexical diversity. Their results showed that the more recently developed lexical diversity measures can explain 74% of variance in human judgement of lexical diversity scores. The findings in their paper showed that human judgement of lexical diversity scores can validate lexical diversity measures. Despite these strengths in their paper, I have highlighted three potential limitations in my critique section: (i) the first limitation relates to the text length problem; (ii) the second limitation relates to the fact that their lemma process has not distinguished polysemous words; and (iii) the third limitation relates to validating human judgement of lexical diversity scores with other writing genres or oral tasks.

2.4 Review of Word Counting Units Studies

The current review section includes two papers concerning word counting units. Considering the word-unit issues flagged in the reviews above, the following section reviews McLean's (2018) and Jarvis and Hashimoto's (2021) papers. McLean (2018) has doubted the use of word family counts as units for L2 language learners, and his paper showed that the

flemma may be a more suitable word counting unit for L1 Japanese participants. Similarly, Jarvis and Hashimoto (2021) investigated how different word counting units affect lexical diversity measures.

2.4.1 McLean, S. (2018): Evidence for the Adoption of the Flemma as an Appropriate Word Counting Unit.

McLean's (2018) article is important because it addressed the need to evaluate word unit choice in second language research and teaching. McLean criticised the common practice of uncritically accepting the use of word families, as established by Bauer and Nation (1993). He argued learners may not be able to recognise all the word family members belonging to the same base word, resulting in an overestimation of learner vocabulary knowledge. McLean reported on an investigation in which he assessed 279 L1 Japanese L2 English language learning participants from three different proficiency levels, as determined by the New Vocabulary Levels Test (NVLT, McLean & Kramer, 2015). He investigated whether the use of the flemma, rather than Bauer and Nation's word families, would provide a better estimate of participant vocabulary knowledge at three different proficiency levels. McLean concluded that the flemma was indeed a more appropriate unit for estimating the vocabulary knowledge of language learners than the word family.

McLean (2018) began by outlining the importance of inflected and derived forms of words in vocabulary learning. He followed Bauer and Nation's (1993) categorisation of word families into seven levels (*each form is a different word* + level 2–level 6). Referring in his paper to a word family as WF6 (Bauer and Nation's affix criteria), McLean highlighted two major issues. The first relates to how Bauer and Nation (1993) defined *word family*:

> *From the point of view of reading, a word family consists of a base word and all its derived and inflected forms that can be understood by a learner without having to*

learn each form separately. (p. 253)

With this claim, they defined the word family from a reading perspective, and the learning of new words implies the learning of new word families. The notion of WF6 knowledge stems from the idea that learners who know one form of the word family are likely to know other forms that belong to the same base word (Webb, 2010).

McLean (2018), however, disagreed with this view, especially for low-proficiency language learners, who may not know the meaning of words that are derived forms of a newly learnt word, such as that knowledge of *marmelise* may not necessarily connote a knowledge of the word *marmelisation*. McLean also contended that because the WF6 standard emanates from an L1 English speaker knowledge standard, it does not follow that English language learners with a different L1, such as Japanese, will share the same knowledge. He argued that the use of the flemma or lemma may thus be more appropriate than WF6.

Because McLean suggested that the lemma or flemma may be more appropriate in assessing L2 language learners than the conventionally accepted WF6 standard, he presented five major studies to support this. Schmitt (2010) stated that the lemma is a better choice of word counting unit than word families. Another four studies (Schmitt & Meara, 1997; Mochizuki & Aizawa, 2000; Ward & Chuenjundaeng, 2009; Sasao & Webb, 2017) have provided evidence that the flemma is a more appropriate word unit than WF6. The participant learners in these studies demonstrated that they have limited knowledge of the inflected and derived forms of word knowledge. Learners with different proficiency levels have shown different levels of knowledge of inflected and derived forms.

Although they provided evidence to support his argument, McLean observed that the four studies are not without their shortcomings. First, there was a lack of reliable data collected in these studies; second, two of the studies (Schmitt & Meara, 1997; Ward &

Chuenjundaeng, 2009) only tested knowledge of suffixes, and did not include validation of prefix assessment; third, the studies do not test knowledge of multiple affixes for the same base word; and fourth, the multiple-choice test format used might have overestimated learners' morphological knowledge.

To investigate whether word families are appropriate word units in estimating word knowledge, McLean called for an assessment of learners' morphological knowledge of WF6 words with multiple affixes. His study attempted to answer the following three questions: First, are there significant differences between L1 Japanese learners' ability to comprehend the base form and their ability to comprehend the inflectional and derivational forms of the same word family; second, are there significant differences between L1 Japanese learners' ability to comprehend the base form and their ability to comprehend the inflected forms of the same word family; and third, can use of the flemma overestimate or underestimate learners' ability to understand the inflected forms and derived forms?

The participants in McLean's investigation were L1 Japanese undergraduates (N=279). They were required to complete the New Vocabulary Levels Test (NVLT, McLean & Kramer, 2015), using a bilingual Japanese version, within 30 minutes. The participants were asked to complete 24 multiple-choice items for each of the first five 1,000-word bands, based on the BNC/COCA (British National Corpus / Corpus of Contemporary American English) word bands. McLean then divided the participants into three groups based on the scores from the NVLT, which included a beginner group (n=85), an intermediate group (n=177), and an advanced group (n=17). An ANOVA (one-way analysis of variance) showed that there were significant differences between the groups ($p<0.001$).

To measure participant vocabulary knowledge of the inflected and derived forms of English words, McLean used a comprehension test which presented high-frequency words from the first 2,000 word families of the British National Corpus (BNC). Twelve words (*use,*

move, collect, center, teach, accept, maintain, develop, standard, circle, adjust, and *publish*) were selected for the test because these words include many inflected and derived forms, according to Bauer and Nation's (1993) word family criteria. For the comprehension test, McLean used 100 sentences in which both inflected forms and derived forms were embedded. The participants were required to translate the underlined L2 English words into their corresponding L1, Japanese, within 30 minutes. The inflected and derived forms of the same base word were presented in the same sets (see the following example sentences for the word *use*, included in the test).

1. *He is useless. =* _____
2. *How do you use this? =* _____
3. *He used the computer yesterday. =* _____
4. *The computer is now usable =* _____
5. *He is using the computer. =* _____
6. *He has used the computer all day. =* _____
7. *He is a user. =* _____
8. *Computers are very useful. =* _____
9. *The usage of this word is common in law. =* _____
10. *Please reuse the paper. =* _____
11. *The bag is reusable. =* _____

Both the NVLT and the sentence comprehension test were scored; multiple raters were used to score the comprehension test, with an inter-rater reliability of over 0.91 using Kappa analysis. In response to each research question, McLean used Cochran's Q test to analyse the data, and treated the base forms, inflected forms, and derived forms as repeated measures. The dichotomous data acquired significant differences between the base form, inflected form, and derived form. For his third research question, McLean used McNemar's

chi-square test to investigate whether the adoption of the flemma was appropriate for his participants.

Regarding his first research question, McLean hypothesised that if there were significant differences found in Cochran's Q analysis, this would show that participants differed in their ability to understand the base forms, the inflected forms, and the derived forms of the words, and so would not support the adoption of WF6 as an appropriate word counting unit. However, if no significant difference were found between the base forms, the inflected forms, and the derived forms, the adoption of WF6 as a word counting unit would be appropriate. Confirming McLean's hypothesis, the Cochran's Q analysis duly indicated a significant difference regarding the number of correct responses to the base word and other members of the same WF6. The large effect size in the study also showed that the participants differed considerably in their ability to understand the base forms and WF6 forms.

In response to his second research question, McLean hypothesised that participants had the same ability to understand the base and inflected forms. The results showed that eight of the twelve tested words (*use, move, collect, teach, accept, maintain, adjust,* and *publish*) showed no significant difference in the flemmas. Only three words (*center, develop*, and *circle*) indicated significant differences, but the effect size was small. These results showed that participants had the same ability to understand the base word and its corresponding inflected forms, indicating that the flemma was an appropriate word unit for the L1 Japanese participants.

In relation to the third research question, the results (see Table 2.26) showed that using the flemma as a word unit underestimated participants' derived knowledge of *-er* for the three tested words *use*, *teach*, and *publish*, but not for *develop* and *adjust*. Regarding inflected knowledge, using flemmas would overestimate the tested words *center* (with *-ed*,

-ing, and *have -ed*), *develop* (with *-ing* adjective), and *circle* (with *-ed*, *-ing*, and *have -ed*). The effect size values were minimal with the existing significant differences and were therefore tolerated. As with the advanced-level participants (n=17), using the flemma as a word unit would not overestimate knowledge but could underestimate 19 derived forms that had been tested.

Table 2.22

The Significance and Effect Size of Differences in the Number of Participants Who Comprehend Base Forms and the Number of Participants Who Comprehend Associated Inflected Forms and Derivational Forms

word form	use	move	collect	center	teach	accept	maintain	develop	standard	circle (verb)	adjust	publish
-ed	0.01	0.0	0.0	0.08*	0.01	0.01	0.01	0.03		0.04*	0.01	0.01
-ing	0.0	0.01	0.02	0.12*	0.01	0.02	0.03	0.03		0.07*	0.02	0.01
-ing adjective								0.09*				
have -ed	0.03	0.02	0.03	0.13*	0.03	0.03	0.03	0.03		0.07*	0.01	0.01
-er	0.02				0.01			0.16*			0.1*	0.01
-able	0.2*		0.2*								0.26*	
-less	0.41*											
…	…	…	…	…	…	…	…	…	…	…	…	…

Note. Effect sizes (Φdisplay differences between the number of participants who comprehend base words and associated inflected or derived forms. *Significant difference. Alpha values for comparisons established by using the Bonferroni adjustment. Adapted from "Evidence for the adoption of the flemma as an appropriate word counting unit," by S. McLean, 2018, *Applied Linguistics*, *39*(6), p. 839. Copyright 2018 Oxford University Press.

McLean's results showed that the participants in his study differed in their ability to comprehend the base forms and WF6, demonstrating that using WF6 as a word counting unit can be inappropriate for measuring language ability. Conversely, using the flemma as the word counting unit only slightly overestimated the beginner and intermediate group participants. McLean, therefore, suggested that the flemma is an appropriate word counting unit for L1 Japanese learners.

McLean highlighted the problems of current vocabulary tests that are based on word family counts such as the Vocabulary Size Test (VST, Nation & Beglar, 2007) and the Vocabulary Levels Test (VLT, Nation, 1983), and he offered an alternative in order to build reliable vocabulary knowledge measures. The vocabulary tests based on word families can overestimate participants' vocabulary knowledge, because learners can guess the meanings of the words when selecting from multiple-choice phrases, even with a limited knowledge of the base form of the words. McLean thus proposed building a more accurate measurement of language learners' vocabulary knowledge in three ways: (i) knowing the participants' inflectional and derivational knowledge levels before conducting vocabulary tests and delivering vocabulary lists; (ii) using derived forms mainly known by participants in the research (i.e., L1 Japanese learners); and (iii) adopting a flemma counting unit as a practical solution for L1 Japanese participants.

Critique

McLean's article offers a valuable contribution to the research relating to inflected and derived vocabulary knowledge. His results showed that participants, especially those at beginner and intermediate levels, have very limited knowledge of derivational forms. To avoid using an inappropriate word counting unit that overestimates participants' morphological knowledge of vocabulary, researchers should be wary of using WF6. McLean

has instead proposed the adoption of the flemma as a practical solution. While his article offers an innovative way of understanding the vocabulary knowledge of L1 Japanese participants, suggesting the flemma as an appropriate word unit to assess word part knowledge, it is not without its weaknesses. These include the number of words selected in the study, the suitability of using the flemma for assessing language skills other than reading, and the fact that McLean only included learners with a single L1, Japanese.

The number of words presented in the study is likely to be problematic. McLean tested 12 English words from the first 2,000 of the BNC on the basis that the low-proficiency Japanese participants could understand these 12 English words. These words had 100 different inflected and derived forms in total, in accordance with the word family levels suggested by Bauer and Nation (1993). Using low-frequency English words (n=12) to judge the morphological knowledge of participants with different language proficiency levels raises concerns, because these 12 English words cannot comprehensively represent participant knowledge of inflected and derived forms of English words. Even the participants identified as being at the same proficiency levels showed discrepancies or individual differences in their inflectional and derivational knowledge of English words belonging to the same or different frequency bands. Including more English words at a wide range of frequency levels would be worth investigating in future studies. Two recent studies conducted by Iwaizumi and Webb (2021, 2022) have suggested that learners' derived vocabulary knowledge is associated with their proficiency levels and vocabulary sizes.

In addition, a problem remains with using the flemma as a word unit for second language research and teaching. McLean's article focused on participants' understanding of morphological knowledge mainly of a single receptive language skill (reading), but not for other language skills (listening, speaking, and writing). Brown et al. (2020) suggested that, for second language studies a smaller word unit, the lemma or flemma, should be adopted

based on a review of the previous morphology studies. The lemma comprises the base word and its inflected forms of the same part of speech (POS) in the English language, including plural, third person singular, present tense, past tense, past participle, -*ing*, comparative, superlative, and possessive forms. The only difference between the lemma and the flemma is that the flemma treats the words in their inflected forms with different POS as the same word. In other words, a flemma as a word unit can include more members than a lemma. However, a lack of sufficient empirical research in this area sheds significant doubt on McLean's suggestion. In addition, morphological knowledge is linked to language proficiency, vocabulary size, and other related factors. Participants with the same proficiency levels may have different derivational knowledge of individual words. Morphological knowledge may to some extent depend on different language skills. The word unit (flemma) recommended in McLean's study for reading may not be suitable for the other language skills (i.e., listening, speaking, and writing). A recent paper (Myint Maw et al., 2022), for instance, suggested that two different word counts (flemma and lemma) might present different interpretations of writing proficiency. What, therefore, needs determining is whether word count units might vary once studies consider the four skills and their assessment.

McLean's article only tested L1 Japanese participants, and we may not necessarily expect the same results for language learners from other language backgrounds. We might see a different set of results if we conducted a replication study for different L1 background populations than those reported in McLean (2018). As Nation and Bauer (2023) stated in their article on morphological awareness:

> *English and many other languages, including Japanese, have words that are made up of meaningful parts and these parts systematically contribute to the meaning of words*. (p. 1)

Crucially, McLean failed to report on participants from other language backgrounds, such as L1 Chinese learners of English. The Chinese and Japanese languages are similar in that there are no singular/plural changes for proper nouns or personal pronouns. However, the two languages differ in that while derivational changes in Japanese are similar to those in English, this is not the case with Chinese.

Were we to use the same method to test L1 Chinese participants as those reported in McLean's article by translating a single word from English to Chinese, it would not detect the morphological changes because no meaningful parts can be added to the changes to the Chinese language, unlike Japanese or English. Thus, testing participants with a wide range of language backgrounds might indicate that no single measure is universally appropriate.

McLean's article contributed significantly to vocabulary research and our understanding of vocabulary learning. It investigated an important issue: the most appropriate word counting unit for language learners that current vocabulary researchers should focus on. The article evaluated L1 Japanese participants from three different proficiency levels (beginner, intermediate, and advanced) by giving them an English-to-Japanese word translation comprehension test that uses sentences for context. The findings showed that L1 Japanese participants had limited knowledge of word families; using word families as a word unit for L1 Japanese participants, would, therefore, overestimate their knowledge of word parts. However, McLean's study indicated that use of a flemma count might only slightly overestimate participant knowledge of the tested words (e.g., *center, circle* as verbs, *develop* with *-ing* forming an adjective). On this basis, McLean suggested that the appropriate word unit is the flemma. We should, though, bear in mind the problems I have highlighted in the evaluation related to suggesting the flemma as an appropriate unit in teaching and research. Other problems with McLean's study include the small number of words tested and the lack of a comprehension test for participants with different L1 backgrounds.

Importantly, though, McLean's article offered possible implications for language learning and teaching. His paper accentuated the need for an emphasis on morphological knowledge because such knowledge relates to learners' vocabulary size and language proficiency level. Research, however, suggests that there is a lack of word-part knowledge training in teaching practice (Dang, 2021). Meanwhile, the way in which word units are processed in research is far from perfect based on current processing tools, and innovation is needed in lexical processing methods (Gablasova & Brezina, 2021). Further research might consider how a focus on morphology that includes derivational forms could be incorporated into language teaching.

2.4.2 *Jarvis, S., & Hashimoto, B. J. (2021): How Operationalisations of Word Types Affect Measures of Lexical Diversity.*

Jarvis and Hashimoto (2021) investigated five different operationalisations of word types within three lexical diversity (LD) measures. The aim was to determine the most helpful LD measures and to demonstrate potential influences of the different word units on each LD index. Their three LD measures consisted of the measures of textual lexical diversity (MTLD), moving average MTLD with wrap-around measurement (MTLD-W), and moving average type-token ratio (MATTR). They employed five different definitions of word types: orthographic forms, lemmas with automated part-of-speech (POS) tags (lemmas-A), lemmas with manually corrected POS tags (lemmas-C), flemmas, and word families. Jarvis and Hashimoto utilised the three LD measures and five types of word units to examine 60 narrative essays written by English, Finnish, and Swedish first-language speakers. Fifty-five human raters evaluated each writing sample, with raters comprising 20 graduate (first-language users of English) and 35 undergraduate students (15 first-language speakers of English; 20 second-language speakers of English with TOEFL iBT scores over 100) studying

linguistics at a university in America. The results for the three LD measures (MTLD, MATTR-50, and MTLD-W) were found to be similar, meaning that it was not possible to determine whether individual measures outperformed others. Mixed results were reported for two of the word units (orthographic forms and lemmas-A) across the LD measures; in contrast, the other three word units (word families, flemmas, and lemmas-C) yielded very similar results across the three LD measures.

Jarvis and Hashimoto presented three main issues in assessing LD: differing operationalisation of types, text length, and human LD ratings. They explained that, in essence, LD relates to word variety in writing and speaking, and word variety can be measured by the number of different words found in written or spoken texts. Tokens are instances of each word occurring in a text, and multiple tokens are repeated items of those found earlier in the text. Conversely, types represent the number of unique words in the text without repetition. Jarvis and Hashimoto stated that most LD measures are variety-repetition (VR) measures, dependent on the counting of types. Since types are so important in LD as measured by VR, Jarvis and Hashimoto contended that it is crucial to reach a theory-and-evidence-based principle of how types should be determined and described in this field.

As mentioned previously, text length has long been a major issue for LD measurement, and one that many studies have questioned. Jarvis and Hashimoto suggested that there are several VR measures that can potentially address this concern. They cited Carroll (1938) as being the first to devise a means to solve the problem of text-length variation, and many other researchers have since followed (e.g., Carroll, 1964; Guiraud, 1954; Herdan, 1960; Johnson, 1939; McCarthy, 2005; McCarthy & Jarvis, 2007; Vidal & Jarvis, 2020; Yule, 1944).

The different measures of lexical diversity are a function of the type-token ratio (TTR, Johnson, 1939), which computes the total number of types (unique words) divided by the

total number of tokens (all words) in a text. The Mean Segmental Type-Token Ratio (MSTTR, Johnson, 1944) divides the text into equal-sized parts and takes the mean of the TTRs of several consecutive samples as the final LD score. However, the problem with MSTTR is that not all the text is used during the calculation process, and this discarding of data has a significant impact on the LD measurements in short texts (McCarthy & Jarvis, 2010).

D (Malvern & Richards, 1997) appears to be the most widely used LD measure. D is calculated using CLAN (Computerised Language ANalysis) software, developed by Brian MacWhinney (2000). The calculation is made through a series of random sampling and curve-fitting procedures by the *vocd* program within CLAN. Jarvis and Hashimoto (2021) affirmed that D increases along with text length, as recorded by both Fergadiotis, Wright, and West (2013) and McCarthy and Jarvis (2007, 2010).

The Measure of Textual Lexical Diversity (MTLD), developed by McCarthy (2005), uses sequential analysis of a sample. A constant TTR value (e.g., under 0.72) is maintained for increasingly longer parts of the sample. For instance, MTLD computes TTR from the first word, the first two words, and so on, until it drops below 0.72. If a TTR value falls below 0.72 at 55 tokens, then the first segment length is 54. The MTLD program then calculates the second segment from token 55, and the final MTLD value is a measure of the mean length of all such segments in which the TTR remains above 0.72.

The Moving Average Type-Token Ratio (MATTR), introduced by Covington and McFall (2010), is also a VR measure. MATTR employs a ‘moving window’, which estimates TTR for each successive window (a fixed length of text, e.g., 50 tokens) until the end of the text; the resultant final MATTR is the mean TTR value of all segments of the text. One advantage of MATTR is that it includes all the words in each text.

MTLD-W, introduced by Kyle, Crossley, and Jarvis (2021) (reviewed in section

2.3.4) and Vidal and Jarvis (2020) (reviewed in section 2.3.3), adopts the moving window approach of MATTR, while also including a 'wrap-around' process that calculates the final segment length by adding words to the initial segment of a text until a TTR of 0.72 is reached. Since MATTR and MTLD appear to be more accurate than other LD indices, and MTLD-W offers improvements on MTLD, Jarvis and Hashimoto (2021) chose to use these three LD measures in their study.

A key issue addressed by Jarvis and Hashimoto (2021) relates to the different ways in which word units can be defined, and how best to operationalise these units in LD studies. Existing possible categories consist of word families, flemmas, lemmas, and orthographic forms. Word families include all derivations and inflections of the same root (Bauer & Nation, 1993). Flemmas cover all inflections of words with the exact spelling, irrespective of the meaning, or part of speech (Pinchbeck, 2014). Lemmas are all the inflections of a word with the same part of speech. With orthographic forms, all inflections are regarded as different types. In Jarvis and Hashimoto's paper, all four different word units were employed.

Jarvis and Hashimoto also considered conceptual elements, both subjective and objective, that comprise the construct of lexical diversity. For this purpose, they referred to Zipf's (1935) study, which regarded lexical diversity as a phenomenon existing fundamentally in the mind, relating more to redundancy in language use (a subjective construct) than to repetition (an objective construct). Similarly, Yule (1944) treated 'lexical richness' as a reflection of the number of types in a learner's mental lexicon. Jarvis (2017) presented a study investigating Zipf's suggestion that human perception of lexical diversity could be superior to other LD measures. In this earlier study, Jarvis observed that human judges were consistent in their ratings without receiving any training. In Jarvis and Hashimoto's (2021) study, which employed the same methodology as Jarvis (2017), the raters appeared to offer high inter-rater reliability (Cronbach's alpha > 0.90), suggesting

again that human raters show excellent agreement without any training or LD rubric. Although few studies have explored the relationships between human ratings of LD and LD measures, Jarvis and Hashimoto posited that VR measures of LD would be able to account to a large degree for the variation in human judgments.

Jarvis and Hashimoto's primary aim was to find out the most effective of the three VR-based LD measurements (MTLD, MATTR, and MTLD-W), when compared with LD as determined by human ratings. An additional goal was to discover which word unit definitions most closely reflect human ratings. Since the application of different word units requires part-of-speech (POS) tagging, Jarvis and Hashimoto wanted to compare the accuracy between automated POS tags and human corrected POS tags. In their study, they ran the cleaned texts through the TreeTagger program automatically, with TreeTagger adding the POS tags and the base form lemma/flemma to the orthographic forms. To establish the accuracy of TreeTagger, they also included the human-corrected POS tag process. Accordingly, their two major research objectives were to determine: (1) which VR measures (MTLD, MTLD-W, MATTR) mirror human ratings; and (2) which word units worked best amongst the three LD measures.

The five categories of word units used in the study were orthographic forms, lemmas-A (lemmas with the automated POS tagger), lemmas-C (lemmas with manually corrected POS), flemmas, and word families.

The corpus data for the study came from Jarvis (2017), with participant English essays written by L1 language users of English (n=13), Finnish (n=31), and Swedish (n=16). Participants were required to write a narrative descriptive essay about an eight-minute-long portion of the Chaplin film *Modern Times*. All writing samples were rated for CEFR writing proficiency by 41 college students majoring in linguistics. Fifty-five human raters judged the LD of each essay, with all raters being undergraduate or graduate students of linguistics. These two groups of human raters did not receive any training, but reportedly had high inter-

rater reliability (Cronbach's alpha=.977 and .983). Jarvis and Hashimoto's data computing process differed from many earlier LD studies. Instead of using existing programs, they created their own Python scripts using the three LD measures (MTLD, MTLD-W, and MATTR). They utilised the TreeTagger program to produce POS tags automatically, and treated lemmas as lemmas-A. They also created a file with corrected POS tags (lemmas-C). Root forms of all words based on Bauer and Nation's (1993) classification of level-six word families were listed.

The results of the research showed high accuracy for TreeTagger (accuracy statistics above 0.90) across major POS divisions except for expletives. Pearson correlations (see Table 2.27) indicated that MTLD had the highest correlations with lemmas-C. The word family worked better with MTLD-W; lemmas-A performed better with MATTR-50.

Table 2.23

Pearson Correlations Between Automated Measures and Mean Lexical Divesity Ratings

	MTLD	MTLD-W	MATTR-50
Orthographic form	0.490	0.411	0.499
Lemma-C	0.528	0.474	0.478
Lemma-A	0.384	0.363	0.501
Flemma	0.516	0.466	0.476
Word family	0.525	0.485	0.485

Note. All coefficients in this table have a *p*-value less than 0.00133. Adapted from "How operationalisations of word types affect measures of lexical diversity," by S. Jarvis, and B. J. Hashimoto, 2021, *International Journal of Learner Corpus Research*, *7*(1), p. 179. Copyright 2021 John Benjamins Publishing Company.

Jarvis and Hashimoto also compared five different operationalisations of word types within the three LD measures through pair-wise comparison. Their results indicated that word family, flemma, lemma-C, and orthographic form outperformed lemma-A in MTLD, and that word family outperformed lemma-A in MTLD-W. Regarding MATTR-50, lemma-A outperformed orthographic form, flemma, and word family, and orthographic form outperformed flemma. Moreover, MATTR-50 outperformed MTLD-W when using lemma-A word types. Linear regression analyses indicated that the highest values were obtained with MTLD by using lemmas-C and word families. In addition, using Cook's distance, Jarvis and Hashimoto investigated the texts that did not meet their designated criteria with either the automated LD measures or different types of operationalisation; five texts were found to be outlier texts.

Jarvis and Hashimoto concluded their paper with a discussion of three main points. First, they retraced their research questions and noted that MTLD correlates most highly with human ratings, followed by MATTR-50 and MTLD-W. They reported no significant differences between LD measures, and their confidence intervals revealed few substantive differences between operationalisations of types following as many as thirty comparisons. However, they suggested that their results do not imply that all the measures or types provide the same function. Their findings also indicated that using the uncorrected POS tags (lemma-A) might lead to unreliable results and that MATTR-50 should not be expected to produce better results with less favourable data. In addition, orthographic forms produced the second strongest correlations with human ratings for MATTR-50 (r=.499), but the second weakest correlations with MTLD (r=.490) and MTLD-W (r=.411). They attributed the reasons to window size, noting that further investigation is necessary into the relationships between measures, window size, and types. In their study, Jarvis and Hashimoto also observed that among types, word families played a constant and significant role, yielding the second

highest correlation with human judgments and MTLD (r=.525), the strongest for MTLD-W (r=.485), and the third highest for MATTR-50 (r=.485). They suggested that this finding was quite unexpected, citing recent papers which claim that lemmas (Kremmel, 2016; Kremmel & Schmitt, 2016) and flemmas (McLean, 2018) are more appropriate than word families to assess vocabulary knowledge. Jarvis and Hashimoto contended that highly professional human raters would judge words with the same root as being less diverse than words belonging to a variety of word families. They concluded that word families, flemmas, and lemmas-C were the three most stable types in their studies.

The second point related to the accuracy of TreeTagger in their study, which was an unexpectedly high 97.2%. In their research, Jarvis and Hashimoto investigated accuracy in connection with three prominent POS tags and they maintained that human examinations of POS mainly focus on these macro level tags. However, they also suggested that even a few POS mistakes can have contrary effects on LD measurements. Therefore, they considered that POS accuracy checking was essential, and is thus something that needs to be implemented in future natural language processing and applied linguistics studies.

Third, Jarvis and Hashimoto suggested that a degree of construct validity was demonstrated in their paper, since each of the three measures accounted for no more than 27.6% of the variance relating to the LD of human judgments, indicating there were factors other than VR measures that could influence LD. Jarvis (2013a, 2013b, 2017) has suggested that there are as many as seven variables that might explain differences in human ratings, and the VR measures of LD under discussion here might only be a small part of the LD construct.

Critique

Jarvis and Hashimoto's (2021) paper represented a pilot study that attempted to validate three VR (variety-repetition) measures of lexical diversity with human rater LD

scores according to five operationalisations of word units. Their study contributed to current LD studies both methodologically and theoretically. The research, however, is not without its shortcomings, so we turn our attention to these in the following sections.

First, the corpus used in the study was problematic. As the authors themselves pointed out, numerous texts in their corpus were short, with some comprising fewer than 150 words. As is widely observed within LD studies, if texts are too short, no differences between the different operationalisations of types among texts can be observed. According to Kyle et al. (2021), Jarvis and Hashimoto (2021), and Vidal and Jarvis (2020), human judgments tend to be influenced by text length, and usually, longer texts receive higher lexical diversity scores because longer texts include a greater range of ideas. The texts in Jarvis and Hashimoto's corpus are all narrative writing samples describing a movie clip, meaning only one genre is covered; furthermore, there are only sixty essays in total. As for the participants, only those at the four proficiency levels from CEFR A1 to B2 level are included in their study, with just two students at B2 level, and nine students at A1. If the intention of the research is to build a standard for current LD studies, then a much larger corpus, which includes more genres and writing samples from participants at different proficiency levels, will be necessary.

Further concerns are that the most appropriate types across all three LD measures have not been determined in the research, and neither have the types that best fit specific measures. Through an examination of three similar LD measures using different definitions of word types, the authors reported mixed results, suggesting that the choice of word unit influences the measurement of LD. Decisions regarding which types are most appropriate for use in future studies need further explication. Jarvis and Hashimoto claimed that the most stable word counting units employed in their study are word families, flemmas, and lemmas-C (lemmas with human corrections). One element that needs considering, however, is that they took word families as being at level six of Bauer and Nation's (1993) levels of word

family. Counting word families in this way reduces learner LD scores during the calculation, because it treats all the words which share the same root as the same type, and so will not distinguish between participants with different levels of word knowledge. In Jarvis and Hashimoto's study, most participants belonged to A2 (n=23) and B1 (n=26) CEFR levels. Lower-level learners will know fewer derivations and inflections, so it is necessary to choose the level of word family carefully, or to consider using other lexical units, in order to gauge their productive vocabulary knowledge.

The third main reservation I have with Jarvis and Hashimoto's paper relates to the human rating of LD scores, and the fact that using numerous human raters to measure LD is hard to implement in practice. In their study, 41 human raters rated both the writing qualities and LD scores, indicating that the same raters had been used twice to rate the same essays. Human raters scored the writing samples using the CEFR Overall Written Production rubric, and they also rated LD after being told that LD is not the same as writing quality. Because the raters did not receive any training in the rating of LD, it is unclear to what extent the CEFR writing rubric might have influenced them. The accuracy of their LD ratings may be questionable.

Regarding the number of human raters in the study, there were 55 reliable raters remaining after four non L1 English raters were removed. To find as many reliable raters as this to rate all the writing samples in a study seems impractical. As mentioned above, Kyle et al. (2021) also adopted direct human judgments in rating all writing samples, but in their paper they used the adjustment scores from two trained human raters until the raters reached an agreement on the same essay. In Kyle et al.'s research, abundance (number of different types) was found to reflect the LD rating most. It should also be pointed out that Jarvis and Hashimoto's study included both L1 English and non-L1 English raters, and the potential influence of the different first languages of the raters has not been considered.

Nevertheless, Jarvis & Hashimoto's paper was important because it employed three widely used LD measures (MTLD, MTLD-W, and MATTR) to evaluate language learner proficiency levels from CEFR A1 to B2. Their study also included five different word units for each LD measurement to investigate how these might influence the LD results in distinguishing different proficiency levels. Their findings indicated that the three LD measures produced different results with each of the five types of word units. These mixed results suggested that lemmas, flemmas, and word families work well with all three LD measurements, with further research required in this area. This evaluation of the study has also highlighted three main weaknesses with Jarvis and Hashimoto's approach. The first relates to the corpus used in their paper: Some of the texts were short (fewer than 150 English words), which undoubtedly influenced the LD scores as judged by the human raters. In addition, the corpus did not include texts written by high proficiency level participants (C1 and C2 learners), and there were only two B2 proficiency level participants. The second shortcoming concerns the failure of the study to determine which word units work better than the others across the three LD measurements. The third weakness of the research relates to the human raters used in the study: fifty-five raters scored the LD, of whom 41 also rated the writing quality of the essays, so whether the raters had been influenced by the CEFR writing rubric remains unclear. Further research is necessary to address these issues emerging from Jarvis and Hashimoto's study. Future studies might look more closely at the construct of lexical diversity, the POS taggers used, and the issues relating to lexical diversity measurements (e.g., the interaction between measures, window size, and operationalisation of types). Research investigating corpora with a wider range of texts is also needed.

2.5 Discussion

The above review of studies has highlighted three key gaps pertinent to vocabulary knowledge tasks and written production: a lack of studies using vocabulary knowledge tasks to predict vocabulary in use or writing proficiency through lexical diversity measures; a lack of development studies investigating vocabulary knowledge development with the same groups of participants through vocabulary tasks and lexical diversity measures; and a lack of studies stating clearly what kinds of word counting units are used for both vocabulary knowledge tasks and lexical diversity measures and applying them with consistency. I have reviewed two main papers addressing vocabulary scores and writing proficiency in sections 2.2 and 2.3. The papers used single vocabulary knowledge tasks or discrete vocabulary knowledge scores to predict writing proficiency. In doing so, I have reviewed a range of vocabulary knowledge tasks in section 2.2 and lexical diversity measures in section 2.3. However, three questions need to be paid special attention to in addressing the relations between vocabulary knowledge tasks and writing proficiency or investigating vocabulary knowledge development by using vocabulary tasks or lexical diversity measures.

1. What types of vocabulary knowledge tasks should be utilised to evaluate participants' vocabulary knowledge? Are there any concerns while using vocabulary tasks?
2. What sorts of measures of lexical diversity ought to be used to predict vocabulary use in writing activities? Are there any concerns while using lexical diversity measures?
3. Should studies keep word count units consistent for responses in vocabulary knowledge tasks and written samples, or can they differ?

Regarding the first question, though vocabulary knowledge can be tested through various vocabulary knowledge tasks, many such vocabulary tasks exist, as reviewed in section 2.2. Should the current study use the existing vocabulary knowledge tasks or develop a new vocabulary knowledge task? As mentioned in my review section regarding vocabulary

knowledge tasks, previously created vocabulary knowledge tasks, such as G_Lex, have not been widely validated in actual vocabulary knowledge assessment contexts. Researchers have widely validated Lex30, but no studies have ever validated Lex30 when used for evaluating writing proficiency. The current dissertation would thus aim to validate existing vocabulary knowledge tasks which have not been widely used in the vocabulary knowledge assessment community.

Laufer and Nation's (1995) study validated the PVLT task with the LFP. Two further studies (Milton et al., 2010; Stæhr, 2008) also validated the VLT task, a receptive vocabulary knowledge task, with writing skills. Treffers-Daller et al.'s (2018) paper explored the relationship between vocabulary scores and writing through discrete vocabulary measures. However, considering writing is a productive skill, there is a lack of studies focusing on validating productive vocabulary knowledge tasks with writing skills. The current dissertation thus aims to focus on investigating productive vocabulary knowledge tasks.

Moreover, as indicated by Fitzpatrick and Clenton (2017), no one vocabulary knowledge task alone can tap all aspects of vocabulary knowledge, and different vocabulary knowledge measures tap different aspects of vocabulary knowledge. The current dissertation will use multiple vocabulary knowledge measures in its ensuing investigations.

As with the second question, as reviewed in section 2.3 about lexical diversity measures, Kyle et al.'s (2021) study emphasised that lexical diversity measures are still under development, and the existing lexical diversity measures cannot capture the whole construct of lexical diversity proposed by Jarvis (2013a, 2013b). Furthermore, no agreement has been established between studies regarding which lexical diversity measures are more effective in predicting writing levels. Considering this, multiple lexical diversity measures should be used, including both previously created and more recently developed measures.

Likewise, text length is another important factor influencing lexical diversity scores,

even though more recently created lexical diversity measures claim not to be influenced by text length (Treffers-Daller, 2013). To solve this problem, Treffers-Daller et al.'s (2018) study selected the middle-200 English words from participants' written production. The current dissertation thus follows Treffers-Daller et al. (2018) and chooses the middle-200 words for all writing samples.

Word counting units are another factor that have been shown to influence vocabulary knowledge task scores and lexical diversity scores (Jarvis & Hashimoto, 2021; McLean, 2018; Treffers-Daller et al., 2018). The traditional word counting unit is the word family. However, McLean's (2018) study proposed that the flemma is a more appropriate word counting unit for L2 English language learners because participants lack the language ability to use word family levels. Considering the participants in my following experimental chapters are mainly from Japan as were McLean's, I use both lemma and flemma as word counting units.

Keeping word counting units consistent across the same study for vocabulary knowledge tasks and lexical diversity measures is crucial because the choice of appropriate word counting unit relates to participants' language proficiency (Nation, 2021). No studies so far have drawn dividing lines between vocabulary size, language proficiency levels, and word family knowledge. As such, I use the lemma as the word counting unit for both responses to vocabulary tasks and writing samples in chapter 3, and I use the flemma as the word counting unit for chapters 4, 5, and 6.

2.6 Conclusion

The papers examined in this literature review chapter highlight a need to conduct empirical studies to address the abovementioned issues. Building on previous studies, I conduct four experiments to examine the role of vocabulary knowledge on written

production. I briefly summarise the contents and research questions of each experimental chapter below.

Chapter 3 is a partial replication of Treffers-Daller et al. (2018) and explores potential relationships between vocabulary measures and L2 written production for participants at CEFR A2 level using lemmas as word counting units. Chapter 3 builds upon Laufer and Nation's (1995) study which explores multiple vocabulary tasks containing both receptive and productive vocabulary knowledge features and lexical diversity measures instead of a frequency-based approach. Chapter 3 presents a first investigation of how four vocabulary tasks can predict IELTS writing levels. The research question asks:

To what extent does a battery of vocabulary tasks predict IELTS writing ability for participants at A2 level?

Chapter 4 focuses on productive vocabulary knowledge tasks and investigates potential relationships between productive vocabulary tasks and L2 written production for participants at CEFR levels B1 to C1. Chapters 4 builds upon McLean (2018) study and Treffers-Daller et al. (2018) study using flemmas as word counting units for both responses in vocabulary tasks and written production. The research question in chapter 4 asks:

To what extent does a battery of productive vocabulary tasks predict IELTS writing ability for participants at levels B1 to C1?

Chapter 5 examines to what extent productive vocabulary tasks can differentiate between IELTS writing scores. I use qualified IELTS raters to judge writing scores for each writing sample. Chapter five builds upon Fitzpatrick and Clenton (2017) study which investigates how vocabulary knowledge tasks can predict language performance for participants with different writing scores. The research questions for chapter five ask:

RQ1: *To what extent can productive vocabulary tasks differentiate between IELTS writing scores for participants at levels B1 to C1?*

RQ2: *Do the results from a comparison of productive vocabulary tasks and lexical diversity measures reflect an increase in writing scores?*

Chapter 6 explores the extent to which productive vocabulary knowledge task scores and lexical diversity measure scores relate over a short study period. Chapter 6 builds upon Fitzpatrick and Clenton (2017) and Vidal and Jarvis (2020) to explore how vocabulary tasks and lexical diversity measures can detect vocabulary knowledge growth. The research questions for chapter 6 ask:

RQ1: *To what extent do productive vocabulary knowledge task scores and lexical diversity measure scores relate to changes over a short study period?*

RQ2: *To what extent do productive vocabulary knowledge task scores and lexical diversity measure scores correlate over a short study period?*

Chapter 3: Exploring Potential Relationships Between Vocabulary Measures and L2 Written Production for A2 Participants: A Partial Replication of Treffers-Daller, Parslow, and Williams (2018)

3.1 Introduction

We often view vocabulary knowledge as essential for language proficiency (e.g., Milton, 2013; Qian & Lin, 2019), and language research often emphasizes the importance of vocabulary knowledge for different language skills. We can see this in research that has shown relationships between vocabulary knowledge and language proficiency in terms of speaking (e.g., Clenton et al., 2020; de Jong et al., 2012; de Jong et al., 2015; Koizumi & In'nami, 2013; Uchihara & Clenton, 2020; Uchihara & Saito, 2016); vocabulary and listening (e.g., Bonk, 2000; Chang, 2007; Stæhr, 2009; Teng, 2016); vocabulary and reading (e.g., Ouellette, 2006; Qian, 1999); and vocabulary and writing (e.g., Laufer & Nation, 1995; Milton et al., 2010; Treffers-Daller et al., 2018).

Such research has indicated that vocabulary knowledge is a key predictor of language proficiency and many testing frameworks have incorporated it, including the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001), a standardised measure of English proficiency used in many European countries. The Council of Europe (2001) describes the position of vocabulary in the CEFR framework as 'major parameters of language acquisition and hence for the assessment of a learner's language proficiency and for the planning of language learning and teaching' (p. 150). Many language proficiency tests measure English language abilities across various CEFR levels (e.g., Business Language Testing Service, Cambridge English, and the International English Language Testing System).

Among the different language proficiency tests, the current chapter emphasises the importance of vocabulary knowledge for different proficiency levels based on the IELTS test.

The IELTS test is an international placement test compatible with the CEFR framework. The IELTS test comprises four main sections, one for each of the four skills: reading, listening, writing, and speaking. Test-takers who possess larger vocabulary knowledge achieve higher IELTS band scores. For example, the British Council (2022) described learners who score in the ninth band, the highest of the bands, as being able to use a wide variety of vocabulary naturally, with sophisticated control of lexical features. The British Council describes those who score in the fifth band as being able to use a limited range of vocabulary, minimally adequate for everyday tasks (British Council, 2022).

The IELTS test includes two types: IELTS Academic and IELTS General Training. They designed the IELTS Academic test for people who wish to pursue a degree in English-speaking countries, whereas the IELTS General Training test is more for everyday English. The writing sections are different for the IELTS Academic and IELTS General Training. The IELTS Academic Writing section includes two writing tasks. Test-takers have to describe graphs, tables, charts, and diagrams. The second task requires test-takers to present their viewpoints and arguments about the topic under discussion in a relatively formal style. Taken together, these two writing tasks account for 25 percent of the total test score. The IELTS General Training writing section also includes two writing tasks. Task one requires a letter to be written in either a formal or informal way. Task two requires essay writing to respond to arguments or problems in a personal style.

The current chapter focuses on the second writing task of the IELTS Academic test since the second task counts double towards the final writing section scores compared to the first task. The second IELTS Academic task requires at least 250 English words, whereas the first requires at least 150 English words. Because we might expect a longer essay to show a greater representation of writing ability, I examine responses to the second writing task rather than the first. Writing ability shows vocabulary knowledge in contextual use. Writing skill

requires learners to use words with a variety of linguistic knowledge (e.g., semantic, morphological, collocational, and syntactic knowledge). The relationship between vocabulary and writing highly depends on linguistic resources, and a clear understanding of vocabulary will allow writers to express themselves accurately and concisely (Schoonen et al., 2011). Thus, language learners must develop their vocabulary to be successful in high-stakes writing assignments (Coxhead, 2012). Learners who take the IELTS test need a quick and highly efficient way to test their vocabulary knowledge to improve their writing levels so that they can begin their higher education degrees. The current study tests participants' vocabulary knowledge and investigates possible relationships between this knowledge and the second IELTS Academic writing task.

Combining these various elements, the experiment reported in this chapter attempts to deepen our understanding of the relationship between writing and vocabulary production. In chapter 2, I observed that the vocabulary score in Treffers-Daller et al. (2018) provided by the PTE Academic does not explain which vocabulary scores can impact the writing score and the CEFR score in their study. Many studies have mentioned that different vocabulary tests tap into different domains of vocabulary knowledge (e.g., Fitzpatrick, 2007; Nation, 2007; Fitzpatrick & Clenton, 2017). Chapelle (2006) has shown that employing multiple vocabulary knowledge measures is necessary to gain more nuances and inferences from the actual performance on vocabulary tasks and to provide a better understanding of vocabulary assessment. The current study, therefore, uses multiple vocabulary measures to assess L2 learners' vocabulary knowledge and various lexical diversity measurements, to evaluate their IELTS written production. Specifically, all participants in this study use one receptive vocabulary task (Vocabulary Levels Test; Nation, 1983; Schmitt et al., 2001) and three productive vocabulary tasks (Lex30; Meara & Fitzpatrick, 2000; G_Lex; Clenton, 2010; Fitzpatrick & Clenton, 2017; and the PVLT; Laufer & Nation, 1995; 1999). Because most L2

studies are based on highly proficient L2 English learners (e.g., Treffers-Daller et al.'s 2018 study examined CEFR B1 to C2 learners), a secondary aim of this study is to explore the vocabulary and writing ability of a group of less proficient learners (CEFR A2). Thus, the research question motivating the current study is:

RQ: *To what extent does a battery of vocabulary tasks predict IELTS writing ability for participants at A2 level?*

3.2 Study

Studies have tended to investigate relationships between writing skills and a single productive vocabulary task score or between writing skills and a single receptive vocabulary knowledge task (e.g., Laufer & Nation, 1995; reviewed in section 2.2.1; Treffers-Daller et al., 2018; reviewed in section 2.3.2). In one such example, Treffers-Daller et al. (2018) demonstrated the relationship between vocabulary score and writing levels and a potential link between vocabulary knowledge, writing ability, and general language proficiency. However, in Treffers-Daller et al.'s (2018) study, the participants had to write an essay on topics set by PTE Academic. The current chapter partially replicates Treffers-Daller et al., requiring participants to complete an IELTS writing task. As a departure from their study, though, participants completed four vocabulary tasks: the VLT, the PVLT, Lex30, and G_Lex.

Table 3.1 shows the descriptive statistics of lexical diversity measures across different proficiency levels in Treffers-Daller et al.'s study. Treffers-Daller et al.'s (2018) study used three different lemmatisation principles. Table 3.1 illustrates their results based on the lemma.

Table 3.1

Basic and Sophisticated Measures of Lexical Diversity Across Different Levels of CEFR (Lemma) in Treffers-Daller et al.'s Study

Measures	B1	B2	C1	C2	Overall means and SD
Types	96.32	104.14	106.32	109.48	103.43 (9.82)
TTR	0.52	0.57	0.58	0.60	0.60 (0.06)
Guiraud	7.09	7.71	7.86	8.08	8.03 (0.74)
D (vocd)	61.88	71.65	73.83	76.61	70.33 (17.28)
HD-D	33.55	34.29	34.55	34.75	34.23 (1.39)
MTLD	58.70	68.52	72.81	77.11	68.37 (17.06)

Note. Reprinted from "Back to basics: How measures of lexical diversity can help discriminate between CEFR levels," by J. Treffers-Daller, P. Parslow, and S. Williams, 2018, *Applied Linguistics*, *39*(3), pp. 315–316 (https://doi.org/10.1093/applin/amw009). Copyright 2018 by the Oxford University Press.

Table 3.2

Correlations Between LD Measures and Pearson Scores in Treffers-Daller et al.'s Study

	TTR	Guiraud	D	HD-D	MTLD	Vocab score	Writing score	Overall score
Types	.973**	.993**	.840**	.843**	.783**	.468**	.447**	.470**
TTR		.993**	.857**	.860**	.787**	.470**	.424**	.455**
Guiraud			.854**	.858**	.790**	.472**	.438**	.466**
D				.925**	.794**	.319**	.290**	.314**
HD-D					.827**	.309**	.276**	.299**
MTLD						.331**	.344**	.338**
Vocab							.765**	.804**
Writing								.920**

Note. Reprinted from "Back to basics: How measures of lexical diversity can help discriminate between CEFR levels," by J. Treffers-Daller, P. Parslow, and S. Williams, 2018, *Applied Linguistics*, *39*(3), p. 318 (https://doi.org/10.1093/applin/amw009). Copyright 2018 by the Oxford University Press. Note. ** p<.01

In Table 3.2, Treffers-Daller et al. reported strong and significant correlations between vocabulary score and writing task score (r=0.765; p<.01), as well as strong and significant correlations between vocabulary score and overall language proficiency level (r=0.804; *p*<.01), and strong and significant correlations between writing score and overall language score (r=0.920; *p*<.01). Taken together, these findings highlighted the strong relationship between vocabulary knowledge, writing ability, and overall proficiency.

Despite the immediate appeal of Treffers-Daller et al. (2018) reporting significant correlations between overall CEFR levels and writing score/vocabulary score, we should use

caution when interpreting their results. Their study's overall CEFR levels stem from combining all four language skills (i.e., listening, speaking, reading, and writing). To investigate their findings, evaluating a single skill (i.e., writing) and using IELTS writing topics and vocabulary tasks might lead to less obfuscation and yield a better and more important relationship. The CEFR levels of participants were influenced by skills that were stronger in contrast to their writing. For instance, some participants may be stronger in reading, listening, and speaking, but weaker in writing. This is because language learners who struggle with the IELTS test are seeking a way to know where their vocabulary knowledge proficiency lies, as well as a test that can actually test their vocabulary knowledge in relation to their writing proficiency.

A further concern with Treffers-Daller et al. (2018) is that they used the Pearson Test of English Academic test (PTE Academic) to determine proficiency, vocabulary knowledge, and writing ability. The PTE Academic is a computer-based English language test adjusted according to CEFR levels. Treffers-Daller et al. (2018) assigned all participants a specific CEFR level based on their performance on 20 assessment items. Accordingly, PTE Academic set the essay topics for participants and their overall scores based on their performance on the test. PTE Academic also provided a vocabulary score based on 15 items and a writing score based on 15 items from PTE Academic itself. What appears clear is that both the vocabulary and writing scores were derived from many discrete variables by the PTE Academic test rather than the actual classroom tests.

On account of the privacy surrounding the PTE Academic test, the extent to which the PTE Academic vocabulary measures reflect vocabulary knowledge in the same or a similar way as the vocabulary measures in the current study remains unknown. Data from the PTE Academic test is not publicly available, so it remains impossible to determine how scores were attributed. The vocabulary score in Treffers-Daller et al. (2018) remains unclear.

Therefore, the reliability and validity of the vocabulary tests in their paper needs further consideration. The current study, accordingly, investigates the relationships between different vocabulary tasks and lexical diversity measures in distinguishing the proficiency levels according to the second IELTS written task.

In the current chapter, I employ three productive vocabulary tasks and one receptive vocabulary task for all participants because vocabulary tasks relate to different aspects of word knowledge, as Fitzpatrick and Clenton (2017) (reviewed in section 2.2.5) suggested. I, therefore, investigated relationships between the different productive vocabulary tasks and writing proficiency in the current study. Part of the purpose of the current study is to investigate the relationships between vocabulary scores and writing scores by employing a multi-task approach and incorporating various vocabulary tasks. These multiple vocabulary tasks investigate participants' vocabulary knowledge and lexical diversity scores within their written production.

3.2.1 Measures

3.2.1.1 Vocabulary Knowledge Tasks. In the current study, productive and receptive vocabulary knowledge tasks investigate potential relationships between vocabulary knowledge and writing abilities among second language (L2) learners. I use three productive vocabulary knowledge tasks for the experiment: Lex30 (Meara & Fitzpatrick, 2000), G_Lex (Clenton, 2010; Fitzpatrick & Clenton, 2017), and the Productive Vocabulary Levels Test (the PVLT; Laufer & Nation, 1999). In the current experiment, I also include one receptive vocabulary knowledge task, the Vocabulary Levels Test (the VLT; Nation, 1983), a receptive version of the PVLT.

A brief review of the three productive vocabulary knowledge tasks is presented first. Lex30, created by Meara and Fitzpatrick (2000), reviewed in section 2.2.2, is a task based on word association and requires participants to write up to four words in response to 30 stimuli.

I gave one mark for each response as per the original scoring criteria. Many studies have validated or used Lex30 (e.g., Baba, 2002; Catala & Espinosa, 2005; Clenton, 2005; Clenton, 2010; Fitzpatrick & Clenton, 2010, 2017; Fitzpatrick & Meara, 2004; González & Píriz, 2016; Uchihara & Saito, 2016; Walters, 2012). G_Lex (e.g., Clenton, 2010; Edmonds et al., 2022; Fitzpatrick & Clenton, 2017), reviewed in section 2.2.5, is a gap-fill task. Learners write up to five English words for each sentence gap. The PVLT (reviewed in 2.2.2), devised by Laufer and Nation (1995, 1999), has been widely used (e.g., Edmonds et al., 2022; Fitzpatrick, 2007; Laufer, 1998; Laufer & Paribakht, 1998; Yamamoto, 2011). As mentioned in section 2.2.2, this test presents one gap for each test sentence at each frequency level.

In addition, one receptive vocabulary knowledge task, the VLT (Nation, 1983; Schmitt et al., 2001), the receptive version of the PVLT, is included in the current experiment. The VLT is a widely used (e.g., Beglar & Hunt, 1999; Laufer, 1998; Laufer & Paribakht, 1998; Schmitt & Meara, 1997; Stæhr, 2008, 2009; Yamamoto, 2011) receptive vocabulary task to access vocabulary knowledge. The VLT is a form-meaning matching task for participants to match the words to the meaning. Participants must write the correct number of words before each explanation, and there are three keys and three distractors for every three words.

3.2.1.2 Lexical Diversity (LD) Measures. Lexical diversity (LD) measures effectively predict language learners' language proficiency levels (Engber, 1995; Treffers-Daller et al., 2018; Treffers-Daller, 2013; Vidal & Jarvis, 2020). Treffers-Daller et al. (2018) described the previously developed LD measures as 'simple' measures, whereas they described the more recently developed LD measures as 'sophisticated'. As a partial replication of their paper, the current study also refers to these LD measures as being either *simple* or *sophisticated*. To extend Treffers-Daller et al.'s. (2018) study, the current study has added several LD measures: namely, Log_TTR, MSTTR, MAAS, MATTR, and MTLD-W. I included these

additional measures to explore the multidimensional features of the lexical diversity construct emphasised by Jarvis (2013a; 2013b). I survey these immediately below.

The current survey presents a summary of both simple and sophisticated LD measures, beginning with simple measures. The simple LD measures presented in the current chapter comprise word types (a simple counting of types), Type-token Ratio (TTR; Johnson, 1944), mean segmental TTR (MSTTR; Johnson, 1944), Log_TTR (Herdan, 1960; sometimes called 'Herdan's C'), and MAAS indices (MAAS, 1972).

TTR (Type-token ratio) is the most widely known measure for capturing the lexical variety in speaking and writing contexts. However, the limitations of TTR are apparent because of its sensitivity to text length. The reason is that learners repeat the vocabulary with the increasing text length. To overcome the limitations of TTR, Johnson (1944) proposed dividing the text into several segments and calculating the average TTR scores for the segments, which he described as the 'mean segment type-token ratio' (MSTTR). The current study includes the mean segmental TTR (MSTTR) measure. I used word type as the word counting unit for each writing sample. Treffers-Daller et al. (2018) found that simple LD measures better predict writing proficiency than sophisticated ones.

Log_TTR, also referred to as Herdan's index *C*, compensates for the text length problems mentioned previously, and the formula of Log_TTR is the number of log types divided by the log tokens. Log_TTR is calculated using the following formula:

$$\text{Herdan's C: } \text{Log\_TTR} = \frac{\log \text{Types}}{\log \text{Tokens}}$$

Root_TTR is known as Guiraud's index (see Guiraud, 1954). Root_TTR shows the ratio between types and the square root of tokens. Numerous papers (e.g., Daller et al., 2013; Daller & Xue, 2007) have shown that Guiraud's index is valid for distinguishing between different proficiency levels. Daller et al. (2013) also suggested that the Advanced Guiraud index is also an adequate measure of lexical sophistication, since it considers the frequency

by removing the first 2K frequency band words because learners are considered to already know these words. Researchers should not take those words into the analysis. In the current chapter, only the Root_TTR score is calculated, and the Advanced Guiraud score is not considered. Root_TTR is calculated using the following formula:

$$\text{Root\_TTR}=\frac{\text{Types}}{\sqrt{\text{Tokens}}}$$

Maas index is an LD measure invented to reduce the text length problem, and the principle of Maas is based on the logarithmic curve. Maas (1972) created the approach. McCarthy and Jarvis (2007, 2010) showed that log correction was effective during the LD correction process. Their research has described the steady MAAS score as text length adjusted to these ranges: 100–154; 154–300; 200–666; and 250–2000. MAAS is calculated using the following formula:

$$\text{Maas index: } a^2=\frac{\text{LogTokens-LogTypes}}{\text{Log}^2\text{Tokens}}$$

The sophisticated LD measures reported in this experiment comprise D (vocd) (Malvern & Richards, 1997; Malvern et al., 2004); HD-D (McCarthy & Jarvis, 2007); and MTLD (McCarthy, 2005). Many papers have widely used and validated the D measure (e.g., Daller et al., 2013; Fergadiotis et al., 2015; Jarvis, 2002; McCarthy & Jarvis, 2007, 2010; Treffers-Daller et al., 2018; Treffers-Daller, 2013), and it needs the software CLAN to compute the score (McWhinney, 2000). The D (vocd) measure estimates a random sampling process of texts, selecting 35 tokens from a random sample of 100 words and then moving from 36 tokens to 50 tokens. Because of the random sampling procedures, CLAN acquires three different D scores, and the final D score is the average of the three D scores. Thus, higher D scores show better LD among writing samples. D is calculated using the following formula:

$$TTR=\frac{D}{N}\left[\left(1+2\frac{N}{D}\right)^{\frac{1}{2}}-1\right]$$

The HD-D measure was first proposed by McCarthy and Jarvis (2007, 2010). They indicated that the D (vocd) index uses the hypergeometric distribution approach to show the token occurrences. They referred to an index based on the hypergeometric distribution as HD-D. Their method in HD-D was choosing a 42-word random sample and then computing the chances that every token can be met in this random sample. The HD-D score is calculated by the chances of the total types appearing within a text.

MTLD is a measure of *textual* lexical diversity proposed by McCarthy (2005), and it is a measure based on textual elements. The author argued researchers should consider both the words and the grammatical (structure) phase. MTLD insists on a fixed TTR value (e.g., 0.72) and computes the TTR from the first word, the first two words, and the adding one word at a time until the TTR falls below 0.72. If, for example, the TTR value falls below 0.72 at 55 tokens, the first segment length is 54. Then the program would calculate the second segment from token 55, and the final MTLD value is the mean length of all these segments.

MTLD Wrap Around (MTLD-W) is an improved measure of MTLD, first mentioned by Vidal and Jarvis (2020). MTLD-W uses the moving window method (the same as MATTR, explained below) and a wrap-around process to compute the final segment by forwarding the last part of a text by adding words from the end to the beginning of the text until it reaches a 0.72 value. MTLD-W is an improved measure of MTLD (Jarvis & Hashimoto, 2021), and it should be more reliable than MTLD.

Covington and McFall (2010) devised the Moving Average type-token ratio (MATTR). It uses a moving window method, such as taking 50 tokens as a segment of a text from the beginning until it reaches the last token of the text. The final MATTR value is the mean value of all segments. The MATTR measure includes all the tokens within each text,

and it also calculates the words along with the textual order, not just choosing words randomly from the text.

3.2.2 Participants

The participants were 29 Chinese undergraduates undertaking their first-year English courses and aged between eighteen and twenty years old. All participants were L1 Chinese speakers of Mandarin. The participants (n=29) had studied English for over 12 years since elementary school. The participants have four English periods for three hours weekly during their first two university years. These classes aim to improve all four English skills: listening, speaking, reading, and writing. Their CEFR levels were A2, as determined by their EFL teachers (they had taken no international English examinations, such as TOEIC, IELTS, or TOEFL). Thus, these participants were considered being of relatively low-level proficiency. The participants in the current chapter gave their consent to take the study, and the process conformed to ethical procedures. All participants joined the experiment voluntarily and reserved the right to withdraw at any time. The Research Ethics Committee of the Graduate School of Humanities and Social Sciences, Hiroshima University, has approved this research (approval number: HR-HUM-000804).

The current study used a priori power analysis (G*Power, Faul et al., 2007) to estimate sample size using a two-tailed test. To determine the minimum sample size, we used Cohen's (1988) guidelines on the effect size of the correlation coefficient. G*Power results show that to achieve power (1-β err prob) equal to 0.8 (80% to detect a difference) for a medium effect (Correlation ρ H1 = 0.3) at a significant level 0.05 (α err prob = 0.05), the minimum required sample size is 84. To detect a large effect size (Correlation ρ H1= 0.5) for correlation coefficient at the significant level 0.05 (α err prob = 0.05), the minimum sample size is 29, resulting in an actual power of 0.81 (81% to detect the significance). The sample

size (n=29) for my current experimental chapter thus meets the requirement for a large effect size.

3.2.3 Methodology

The participants were required to complete three productive vocabulary knowledge tasks (Lex30, G_Lex, the PVLT), one receptive vocabulary knowledge task (the VLT), and one IELTS writing task (see Appendix A). Participants had to complete all tasks during two class periods within one week, which meant that they first completed Lex30, G_Lex, and the VLT tasks during their first class time and then the PVLT task and writing task the following week. To begin the process, I gave the participants a very general IELTS writing topic, and the purpose was to give them more opportunities to elicit more words during their writing process. Prior to data collection, participants received a brief explanation of the project. They completed all tasks within the class time within the two weeks. I used pen and paper for data collection, and the instructors controlled the time for each task. The original task instruction specified the timing for the tasks: it instructed participants to complete both Lex30 and G_Lex within 15 minutes, and the VLT and the PVLT tasks within 25 minutes. I gave learners 40 minutes to complete the writing task. The experiment was conducted in March 2019.

3.2.4 Data Analysis

I analysed the data by having all paper documents, including Lex30, G_Lex, and the PVLT, and writing samples converted into electronic data by experienced research assistants.

3.2.4.1 Vocabulary Tasks Data Analysis. I corrected all spelling mistakes for the productive vocabulary knowledge tasks from the participants. The Lex30 and G_Lex tasks were lemmatised according to the lemma criteria proposed by Meara and Fitzpatrick (2000). For the PVLT, because of the unique characteristics of the stimulating words in the PVLT, there was no need to conduct the lemmatisation process. The PVLT task requires participants

to complete the predetermined words for five different frequency levels (2K, 3K, 5K, UWL, and 10K), with the first few letters of the words being given for each gap. The current study uses the lemma as a word unit, and it is impossible to find the lemma words in the PVLT because the predetermined words in the PVLT were selected based on word families.

I profiled all three productive vocabulary knowledge tasks using AntWordProfiler (Anthony, 2022) to divide responses based on different frequency levels. The BNC/COCA word lists, created by Nation (2017), were imported into this program. The AntWordProfiler sorted words through frequency bands, and the output comprised word types, not tokens. Type counts should provide a means of objective evaluation of word knowledge, meaning that all repeated tokens were treated as the same words. I treated the calculation of vocabulary tasks the same as Fitzpatrick and Clenton (2017). Specifically, I processed word knowledge items exceeding the 1K frequency. The final analysis for the three productive vocabulary knowledge tasks counted only the percentage score of word types. I specify this because it is unknown whether Treffers-Daller et al. (2018) used percentage scores or raw scores in reporting vocabulary scores.

In the current study, I scored the Lex30 and G_Lex tasks by excluding all words produced in the 1K band, which means the participants' vocabulary knowledge is calculated from 2K and above, as in Fitzpatrick and Clenton (2017). I removed these words because the PVLT and the VLT tasks do not include the 1K level, so removing the 1K level from Lex30 and G_Lex ensures consistency in frequency levels across all tasks.

After removing all vocabulary knowledge belonging to the 1K level, I calculated the raw scores of all vocabulary tasks by giving one mark to each correct response. Subsequently, I converted all raw scores into percentage scores because the maximum score for each vocabulary task was different. For Lex30 and G_Lex, participants had to write up to 120 English words. The PVLT task required participants to complete a predetermined 90 words in

total. The VLT required learners to match the explanations with 150 words. I computed the VLT task scores by counting the corrected matching for all frequency levels. Since the VLT task starts from the 2K level, I gave one point to all corrected answers from the participants in each frequency level. I computed the VLT task scores by dividing all the correct responses by 150. In the current chapter, I only consider participant percentage scores. I am interested in determining the vocabulary knowledge of each participant; however, the approximate calculation of a number cannot represent the level of knowledge of the participants. To calculate the raw scores, each word produced beyond the 1K level is given a point, which are then added together. I calculated percentage scores by dividing the raw scores by the total number of words (n=120) in Lex30 and G_Lex. For all vocabulary tasks, I presented only percentage scores exceeding 1K in the following tables, similar to what previous studies have done (e.g., Clenton, 2010; Fitzpatrick & Clenton, 2010; 2017).

3.2.4.2 Writing Samples Data Analysis. I treated all writing samples in the same way as Treffers-Daller et al. (2018), with the main difference being that I treated the data in the current study manually and used an automated Python script (Treffers-Daller et al. (2018) used CLAN to clean their writing samples). I corrected spelling mistakes to prevent the software from counting the words as different types. Before computing the lexical diversity measures, I needed to clean the data. I deleted proper names, such as the names of cities, trademarks, names of people, and local food names. Further, I amended abbreviations such as *TV* to *television*; I removed numbers written in figures such as *1990* and *50*, but retained numbers expressed as *fifty* in the sample.

I kept the text length constant using a Python script and selected only the middle 200 English words as in Treffers-Daller et al. (2018). I removed some written samples because they were less than 200 English words. The main reason I only selected the middle 200 English words was to be consistent in addressing an unresolved issue with lexical diversity

measures since LD scores vary with different text lengths. As text length grows, LD scores in writing become lower.

Selecting the appropriate word units remains a significant issue in this field because the lexical unit is a critical element for LD scores (e.g., Jarvis & Hashimoto, 2021). Treffers-Daller (2013) used lemmas to predict the language abilities of their group of French L2 English learners because of the inflected features of the French language. Treffers-Daller et al. (2018) used three different lemmatisation standards to analyse their writing samples. Their results showed the lemma was more accurate in predicting participants' writing abilities than no lemmatisation and word families. Since the current study is a partial replication of Treffers-Daller et al. (2018), I also used lemmas (Nation, 2016) as the word unit for the analysis of all IELTS writing samples. Myint Maw et al. (2022) concluded that both lemma and flemma offer a different means of predicting writing proficiency compared to simple word counts (tokens). Since the level of the participants in the experiment reported in this current study (CEFR=A2) comprises basic English language users, and because they probably lacked the morphological knowledge of English words, I used a lemmatisation process to calculate the LD indices for all writing samples.

To calculate the LD values, I computed D (vocd) scores using CLAN (MacWhinney, 2000) and the remaining LD scores by TAALED, a text processing tool to calculate various LD measures (Kyle et al., 2021). To mirror Treffers-Daller et al.'s study, I scored with CLAN for the D (vocd) score. However, in a departure from Treffers-Daller et al. (2018) study, I calculate the rest of the LD measures by TAALED, whereas Treffers-Daller et al. used SPSS for the Guiraud index (also known as Root_TTR), excel spreadsheets for HD-D, and Gramulator for MTLD. I am doing this because I want to lemmatise all the writing samples during data processing. A built-in command in CLAN for D (vocd) can calculate the D score.

Similarly, the TAALED software automatically distinguishes homographs based on internal part-of-speech tags and calculates all LD indices from its lemma forms.

To mirror Treffers-Daller et al. (2018), all the writing samples needed to be converted into a CHAT format to adapt to the CLAN program. However, as a departure from their study, I used the morphosyntactic tier (the analysis interface presented in CLAN for the lemmatised words) command for writing samples and computed the D (vocd) scores only based on the mor tier. CLAN can calculate the D (vocd) measure through the lemma with the vocd command: vocd +sm;*,o% @. I converted all writing samples to a *.txt* format before utilising TAALED to calculate LD scores.

3.3 Results

The research question for the current chapter asked: *To what extent does a battery of vocabulary tests predict IELTS writing ability for participants at A2 level?* This section reports the vocabulary task scores, lexical diversity measures results, and their predictability to lexical diversity scores.

3.3.1 Vocabulary Task Results

Table 3.3 shows the descriptive statistics of the productive vocabulary knowledge percentage scores. The results in Table 3.3 show that the mean scores for Lex30 (mean=12.44%), G_Lex (mean=12.64%), and the PVLT (mean=39.81%) vary. The Lex30 (mean=12.44) and G_Lex (mean=12.64) mean scores are similar, and the PVLT task has the highest mean score among the three vocabulary tests.

Table 3.3

Productive Vocabulary Knowledge (PVK) Tasks Descriptive Statistics

PVK measures (n=29)	Minimum	Maximum	Mean	SD
Lex30%	0.83	26.67	12.44	6.28
G_Lex%	3.33	26.67	12.64	7.43
PVLT%	1.11	58.89	39.81	16.62

The current chapter uses the Shapiro-Wilk test to determine whether the data set meets the assumption of normal distribution. Results show that the significant values of G_Lex (p=0.02), the PVLT (p=0.002), the VLT (p=0.000), and MAAS (p=0.000) violate the normal distribution ($p<0.05$). For the data that has violated normal distributions, I ran Spearman's rho correlations and robust regressions using bootstrapping.

Table 3.4 shows the correlations between the three productive vocabulary knowledge task scores. I calculated correlations between the tasks. The results in Table 3.4 show no significant correlations between the PVLT and G_Lex, nor between the PVLT and Lex30, but a moderately significant correlation between Lex30 and G_Lex (r_s=.528**, $p<0.01$).

Table 3.4

Correlations Between Productive Vocabulary Knowledge (PVK) Tasks

PVK measures (n=29)	G_Lex%	PVLT%
Lex30%	.528**	-.221
G_Lex%		.074

Note. ** Significant at the 0.01 level (2-tailed)

Table 3.5 shows the correlations between the VLT and the three productive vocabulary knowledge tasks. As seen in Table 3.5, there are no significant correlations

between the VLT and Lex30 (r_s=.017, p>0.01), nor between the VLT and G_Lex (r_s=.234, p>0.01) in Table 3.5, nor between the VLT and the PVLT (r_s=.347, p>0.01).

Table 3.5

Correlations Between Productive Vocabulary Knowledge (PVK) Scores and Vocabulary Levels Test (VLT) Scores

PVK scores (n=29)	Lex30%	G_Lex%	PVLT%
VLT scores	.017	.234	.347

Note. ** Significant at the 0.01 level (2-tailed)

3.3.2 Lexical Diversity Measure Results

Table 3.6 shows the lexical diversity measures' descriptive statistics. A high LD score indicates a high level of writing ability. The results in Table 3.6 illustrate the mean scores and the SD for the LD measures. The mean scores of LD measures show that the HD-D (mean=.79), MSTTR (mean=.75), and MATTR (mean=.75) are very similar, even though they use different formulas. The MTLD score (mean=62.24) and MTLD_W score (mean=60.58) remain slightly different. D (vocd) (15.66), MTLD (SD=16.63), and MTLD-W (SD=16.26) are much higher than the other LD measures, whereas the simple measures, such as Log_TTR (SD=.01) and MAAS (SD=.01) have the lowest SD score. The high SD values show that D (vocd), MTLD, and MTLD-W values are far from the mean values. The lowest SD with Log_TTR and MAAS scores demonstrate their values are clustered close to the mean values.

Table 3.6

Descriptive Statistics of Lexical Diversity (LD) Measures

LD measure	Minimum	Maximum	Mean	SD
Types	89	125	108.41	8.87
TTR	.45	.63	.55	.05
Root_TTR	6.29	8.84	7.72	.63
Log_TTR	.85	.91	.89	.01
MSTTR	.68	.83	.75	.04
MAAS	.04	.07	.05	.01
D (*vocd*)	44.91	109.73	69.8	15.66
HD-D	.73	.86	.79	.03
MTLD	36.95	101.29	62.24	16.63
MTLD_W	36.50	98.41	60.58	16.26
MATTR	.66	.82	.75	.04

Table 3.7 shows the correlations between lexical diversity (LD) measures. Since it normally distributed the values for LD measures except for the MAAS scores, I used Pearson's r and Spearman's rho correlation analysis. The LD measures appear highly correlated, as Table 3.7 shows, and such strong correlations indicate that different LD measures are assessing the same construct; these measures developed so far are based on an adjustment of types and tokens (Treffers-Daller, 2013; Jarvis & Hashimoto, 2021). Because of its log function proposed by MAAS (1972), an increase in types results in lower MAAS values for the MAAS measure (see section 3.2.1.2 for the MAAS formula). Thus, low MAAS scores equate to high LD, and high MAAS scores equate to low LD scores. Therefore, the

strong negative significant correlations shown in Table 3.7 are actually significant positive correlations between MAAS scores and the other LD measures.

Table 3.7

Correlations Between Lexical Diversity (LD) Measures

LD measure	TTR	Root_TTR	Log_TTR	MSTTR	MAAS	D (vocd)	HD-D	MTLD	MTLD_W	MATTR
Types	.988**	.996**	.967**	.733**	-.916**	.766**	.839**	.848**	.888**	.780**
TTR		.997**	.977**	.739**	-.919**	.768**	.835**	.859**	.888**	.791**
Root_TTR			.978**	.734**	-.915**	.768**	.840**	.856**	.888**	.784**
Log_TTR				.721**	-.921**	.755**	.832**	.840**	.875**	.773**
MSTTR					-.699**	.807**	.851**	.890**	.889**	.927**
MAAS						-.734**	-.794**	-.795**	-.843**	-.771**
D (vocd)							.969**	.873**	.924**	.829**
HD-D								.919**	.950**	.869**
MTLD									.951**	.888**
MTLD_W										.890**

Note. ** Significant at the 0.01 level (2-tailed)

3.3.3 The Results Between Vocabulary Knowledge Tasks and LD Measures

A primary aim of the current chapter was to explore potential relationships between the various vocabulary tasks and lexical diversity measures. Table 3.8 shows the comparisons between the productive task scores and LD measures. Specifically, Table 3.8 shows the correlations between four vocabulary tasks (Lex30, G_Lex, PVLT, and VLT) and a spectrum of major LD measures. Table 3.8 demonstrates no significant correlations between the three productive vocabulary knowledge scores and the various LD scores. Table 3.8 also reports no significant correlations between LD measures and vocabulary knowledge task scores. This means that the LD scores in writing do not correlate or move in sync with vocabulary knowledge task scores. When vocabulary task scores increase, the LD scores will decrease.

Table 3.8

Correlations Between Vocabulary Tasks Scores and Lexical Diversity (LD) Scores

LD measures	Lex30%	G_Lex%	PVLT%	VLT%
Types	.030	.086	-.225	-.133
TTR	.070	.082	-.221	-.116
Root_TTR	.059	.076	-.226	-.115
Log_TTR	.093	.063	-.260	-.147
MSTTR	.066	.116	-.305	-.129
MAAS	-.096	-.042	.218	.096
D (*vocd*)	.152	.091	-.268	-.086
HD-D	.147	.114	-.232	-.100
MTLD	.048	.049	-.377*	-.204
MTLD_W	.070	.059	-.261	-.164
MATTR	.195	.102	-.298	-.193

Note. ** Significant at the 0.05 level (2-tailed)

To examine whether the vocabulary knowledge tasks can predict LD scores, I ran a robust, simple standard linear regression analysis using bootstrapping (Larson-Hall, 2015). The results in Table 3.9 show that four vocabulary knowledge tasks can explain minor variance in lexical diversity scores. The biggest explanation was between VLT and MATTR (R^2=0.095), followed by the PVLT and MSTTR (R^2=0.075), and then the PVLT and MTLD (R^2=0.069).

Table 3.9

Regression Analyses Between Vocabulary Tasks Scores and Lexical Diversity (LD) Scores

variable	R^2	sr^2	Intercept	B	95% Confidence Interval	
					Lower Bound	Upper Bound
Lex30→Type	0.001	0.030	107.879	0.043	-0.515	0.601
Lex30→TTR	0.005	0.070	0.544	0.001	-0.002	0.003
Lex30→Root_TTR	0.003	0.059	7.644	0.006	-0.034	0.046
Lex30→Log_TTR	0.009	0.093	0.884	0.000	-0.001	0.001
Lex30→MSTTR	0.004	0.066	0.745	0.000	-0.002	0.003
Lex30→MAAS	0.014	-0.116	0.051	0.000	-0.001	0.000
Lex30→D (vocd)	0.023	0.152	65.079	0.379	-0.594	1.353
Lex30→HD-D	0.022	0.147	0.782	0.001	-0.001	0.003
Lex30→MTLD	0.002	0.048	60.671	0.126	-0.919	1.171
Lex30→MTLD_W	0.005	0.070	58.324	0.181	-0.839	1.201
Lex30→MATTR	0.038	0.195	0.737	0.001	-0.001	0.004
G_Lex→Type	0.001	-0.025	108.794	-0.030	-0.501	0.441
G_Lex→TTR	0.000	-0.009	0.551	0.000	-0.003	0.002
G_Lex→Root_TTR	0.001	-0.023	7.743	-0.002	-0.036	0.032
G_Lex→Log_TTR	0.002	-0.041	0.888	0.000	-0.001	0.001
G_Lex→MSTTR	0.002	0.039	0.747	0.000	-0.002	0.002
G_Lex→MAAS	0.001	0.026	0.049	0.000	0.000	0.000
G_Lex→D (vocd)	0.004	0.064	68.101	0.134	-0.696	0.965
G_Lex→HD-D	0.001	0.035	0.789	0.000	-0.001	0.002
G_Lex→MTLD	0.002	-0.045	63.522	-0.101	-0.984	0.782
G_Lex→MTLD_W	0.000	-0.006	60.742	-0.013	-0.877	0.851
G_Lex→MATTR	0.001	0.025	0.750	0.000	-0.002	0.002
PVLT→Type	0.010	-0.100	110.539	-0.053	-0.263	0.156
PVLT→TTR	0.010	-0.098	0.561	0.000	-0.001	0.001
PVLT→Root_TTR	0.011	-0.105	7.878	-0.004	-0.019	0.011
PVLT→Log_TTR	0.025	-0.158	0.892	0.000	0.000	0.000
PVLT→MSTTR	0.075	-0.273	0.778	-0.001	-0.002	0.000
PVLT→MAAS	0.011	0.104	0.048	0.000	0.000	0.000
PVLT→D (vocd)	0.051	-0.225	78.255	-0.212	-0.575	0.150
PVLT→HD-D	0.044	-0.210	0.806	0.000	-0.001	0.000
PVLT→MTLD	0.069	-0.263	72.700	-0.263	-0.644	0.118
PVLT→MTLD_W	0.059	-0.243	70.039	-0.238	-0.612	0.137
PVLT→MATTR	0.049	-0.221	0.772	-0.001	-0.001	0.000
VLT→Type	0.013	-0.114	111.681	-0.043	-0.191	0.105
VLT→TTR	0.017	-0.130	0.570	0.000	-0.001	0.001
VLT→Root_TTR	0.016	-0.126	7.975	-0.003	-0.014	0.007
VLT→Log_TTR	0.044	-0.211	0.897	0.000	0.000	0.000

VLT→MSTTR	0.045	-0.212	0.779	0.000	-0.001	0.000
VLT→MAAS	0.033	0.182	0.045	0.000	0.000	0.000
VLT→D (vocd)	0.022	-0.149	77.335	-0.099	-0.359	0.160
VLT→HD-D	0.020	-0.141	0.805	0.000	-0.001	0.000
VLT→MTLD	0.054	-0.232	74.687	-0.164	-0.435	0.107
VLT→MTLD_W	0.051	-0.227	72.444	-0.157	-0.422	0.109
VLT→MATTR	0.095	-0.309	0.790	0.000	-0.001	0.000

Note. Bootstrap results are based on 2000 bootstrap samples.

3.4 Discussion

The purpose of the current study is to investigate the extent to which vocabulary tasks predict writing proficiency in a partial replication of Treffers-Daller et al. (2018). The research question for the current replication chapter was: *To what extent does a battery of vocabulary tests predict IELTS writing ability for participants at A2 level?* The battery of measures comprised one receptive vocabulary test (VLT) and three productive vocabulary tests (Lex30, G_Lex, and PVLT). The current chapter is a partial replication of Treffers-Daller et al. (2018), who reported the descriptive statistics of LD measures based on the lemma principle (Table 3.1). The results reported in the current study differ from Treffers-Daller et al.'s (2018) study, in terms of both descriptive statistics and correlations. The following discussion briefly outlines these differences, beginning with the focus of the current chapter: the different vocabulary scores.

The current study's vocabulary task scores differ from Treffers-Daller et al.'s (2018) study. Treffers-Daller et al. (2018) provided their vocabulary score by using the PTE Academic test, which remained confidential, and they did not report the actual vocabulary values. The current chapter includes four discrete vocabulary tasks: the VLT, the PVLT, Lex30, and G_Lex. The results showed significant correlations between Lex30 and G_Lex scores (r_s=.528**, see Table 3.4). The significant correlations between vocabulary scores are similar to those reported in previous research (e.g., Edmonds et al., 2022; Fitzpatrick & Clenton, 2017; Walters, 2012).

The differences between the current study and Trefffers-Daller et al. (2018) in descriptive statistics may result from participants' English proficiency levels. Participants with higher proficiency have a greater vocabulary knowledge than participants with lower proficiency. The proficiency levels of participants in Treffers-Daller et al. (2018) study were from B1 to C2, while the level for the current study's participants was A2. Participants in Treffers-Daller et al. (2018) had relatively higher mean LD scores (Table 3.1) than the participants in the current study (see Table 3.6).

The current study also differs from Treffers-Daller et al.'s (2018) study regarding the correlation between vocabulary scores and LD measures. In the current chapter, I did not find significant correlations between the receptive vocabulary task (VLT) and lexical diversity (LD) measures (Table 3.8). There were also no significant correlations between the three productive vocabulary tasks and the LD measures (Table 3.8), although minor variance in lexical diversity measures could be explaned by vocabulary knowledge tasks (Table 3.9). Treffers-Daller et al. (2018) reported significant correlations between vocabulary knowledge scores and lexical diversity measures (see Table 3.2).

Another difference between the current study and Treffers-Daller et al. (2018) is that the vocabulary tasks and LD measures used differ. First, the current study uses IELTS writing samples, whereas Treffers-Daller et al. (2018) used the writing data from the PTE Academic. Second, the current chapter uses four vocabulary tasks; Treffers-Daller et al.'s study did not use these tasks. Third, I added several LD measures to the current dissertation as compared to Treffers-Daller et al. (2018) by incorporating a comprehensive combination of simple LD measures (Log_TTR; MSTTR; MAAS) and sophisticated LD measures (MTLD_W; MATTR). Adding more LD measures is justified because, based on combining various LD measures, there is a greater chance to capture more features of Jarvis' (2013a, 2013b) LD construct.

The main issue relates to the proficiency levels of the participants in the current study. With participants at the A2 level, it is possible that their vocabulary scores (see Table 3.3) do not represent their vocabulary use in their IELTS writing. If this is true, the elicited words in the tasks, particularly the productive vocabulary tasks, do not accurately represent their actual vocabulary knowledge in IELTS writing. One explanation is that writing entails complex and comprehensive lexical knowledge. Low-level participants have limited vocabulary knowledge and lack the ability to put their limited vocabulary knowledge into their IELTS writing. To remedy this, the next chapter conducts another study which includes participants with higher proficiency levels to better mirror Treffers-Daller et al.'s findings.

3.4.1 Limitations

The limitations of the current study relate to the levels of participants. Treffers-Daller et al.'s (2018) study included four different participant groups ranging in proficiency level from B1 to C2, while the current study limited the participants to an A2 proficiency level. It would be interesting to explore how participants at the specific level (A2) would perform in the study by Treffers-Daller et al. (2018). However, I will address this concern by investigating a broader range of participants to be evaluated in the experiment reported in chapter 4, which follows.

3.4.2 Conclusion

The current chapter presents a first perspective on how IELTS writing levels can be evaluated using receptive and productive vocabulary tests. I included four vocabulary tests in the current study: one receptive vocabulary test (VLT) and three productive vocabulary tests (Lex30, G_Lex, and PVLT). The results show no significant correlation between these four vocabulary tests and IELTS writing scores.

Since high-level language learners possess a more extensive vocabulary knowledge than low-level language learners, the next chapter (4) will include more participants of

different proficiency levels. Since writing is a productive skill, it is likely that productive vocabulary tests correlate significantly with IELTS writing. The following chapters will focus on productive vocabulary tests.

Chapter 4: Exploring Potential Relationships Between Productive Vocabulary Tasks and L2 Written Production for B1 to C1 Participants

4.1 Introduction

The discrepancies in results between the experiment reported in chapter 3 and those reported in Treffers-Daller et al. (2018) identified three major issues: differences in vocabulary tests and writing, absence of significant correlations between vocabulary tasks and lexical diversity measures, and the relation of word units to vocabulary task scores and writing samples. Section 3.4 firstly emphasised that the vocabulary tasks in the experiment reported in chapter 3 differed from the vocabulary scores in Treffers-Daller et al. (2018). The experiment reported in chapter 3 used different writing samples compared with Treffers-Daller et al. (2018). Second, chapter 3 found no significant correlations between the vocabulary tasks (the VLT, the PVLT, Lex30, and G_Lex) and lexical diversity measures and only minor explanations of lexical diversity scores. I concluded that this lack of correlations and significant explanations was because of the proficiency levels of the participants (CEFR=A2) compared with Treffers-Daller et al. (2018), in which the proficiency of participants ranged from B1 to C1 level. Third, Treffers-Daller et al. (2018) used three lemmatisation standards for lexical diversity measures. They adopted no lemma (simple count of types), lemma, and lemma 2 (word families to level 3, based on Bauer & Nation, 1993). In Treffers-Daller et al.'s study, they did not count the flemma as a word unit for the LD measures.

The current chapter, therefore, uses flemma (Nation, 2016) as a word unit for the IELTS writing samples when I compute the scores of lexical diversity measures. I analyzed three vocabulary tasks (Lex30, G_Lex, and PVLT) using flemma before calculating the vocabulary scores to keep the word units consistent. Chapter 3 reported the lemma as the word unit for vocabulary task scores, as in previous studies (e.g., Meara & Fitzpatrick, 2000;

Fitzpatrick & Clenton, 2017) and the lemma as the word unit for the lexical diversity measures. However, because I reported morphological knowledge can both influence vocabulary task scores (Brown et al., 2020; McLean, 2018) and lexical diversity scores (Jarvis & Hashimoto, 2021, section 2.4.2), it is important to determine the appropriate word unit for the current study. Brown et al. (2020) concluded that the most appropriate word units could be the lemma or flemma depending on learner proficiency. The current study thus employs the flemma as a lexical unit before calculating LD scores. Jarvis and Hashimoto (2021) proposed that three sophisticated lexical diversity measures (MTLD; MTLD_W; MATTR) might better predict writing scores with three different word units (lemmas, flemmas, and word families). McLean's (2018) paper suggested that a flemma count was a more appropriate word counting unit for EFL learners. However, McLean's study also stated the belief that the participants in his study did not have the language ability to understand derivational forms. Because the participants in the current study are from Japan and China, and their CEFR levels vary from B1 to C1, I assume the participants already have the word knowledge of parts-of-speech (POS).

The current chapter continues by exploring the relationships between vocabulary tasks and IELTS written production with L2 English learners using the flemma as the word unit for both vocabulary tasks and LD measures. Kyle (2019) also criticised the fact that many tools have appeared claiming to lemmatise text, but in fact, they were flemmatising them. In addition, it is rare to find that learners will use homographs in their texts if only the middle 200 English words are chosen from the texts for analysis. Therefore, in the current study, following McLean (2018), I hypothesise the flemma is an appropriate word unit for both vocabulary tasks and IELTS writing samples.

Building on chapter 3, the current chapter explores the same question as chapter 3 but with higher proficiency level participants. The current chapter focuses on productive

vocabulary tasks by including the same three productive vocabulary tasks: the PVLT, Lex30, and G_Lex. In the experiment reported in the current chapter, I have not included the receptive vocabulary task (the VLT) since the results reported in chapter 3 suggest we cannot use the receptive vocabulary task as a predictor of writing ability. A recent paper (Edmonds et al., 2022) questioned that the PVLT might not be 'the best choice for concurrent validity studies concerning the assessment of productive vocabulary knowledge' (p. 8). They showed that the PVLT 'patterns with the measure of receptive vocabulary knowledge (the VLT)' (p. 8). They interpreted the performance of the PVLT in their study 'as representing receptive vocabulary knowledge' (p. 9). Ironically, despite its name, the PVLT might also be an indicator of receptive vocabulary knowledge, so, again, there may be no need for the VLT in the current experiment.

Moreover, IELTS writing tasks belong to the area of productive skills for English language learners. I want to limit the number of vocabulary tasks by focusing on productive vocabulary tasks with participants of higher language proficiency levels than those reported in chapter 3. On the other hand, previous studies (e.g., Laufer & Nation, 1995, section 2.2.1; Treffers-Daller et al., 2018) have investigated potential relations between vocabulary knowledge and writing production, focusing on either a single productive vocabulary task (Laufer & Nation, 1999) or receptive vocabulary scores (Treffers-Daller et al., 2018). Accordingly, the research question for the current chapter is:

RQ: *To what extent does a battery of productive vocabulary tasks predict IELTS writing ability for participants at levels B1 to C1?*

4.2 Study

To investigate whether productive vocabulary tasks can predict IELTS writing levels, the experiment reported in the current chapter uses three productive vocabulary tasks (Lex30, G_Lex, and the PVLT) for all participants. The levels of participants in the current chapter

are higher than those reported in chapter 3. In addition, previous papers (Meara & Fitzpatrick, 2000; Fitzpatrick & Clenton, 2017) employ lemmas as word units for Lex30 and G_Lex for English language learners with different proficiency levels. Meanwhile, the lemma standard for Lex30 in Meara and Fitzpatrick's (2000) paper is based on the level 2 (inflected suffixes) and level 3 (most frequent affixes) criteria proposed by Bauer and Nation (1993, pp. 29–30). This lemma standard (Meara & Fitzpatrick, 2000) in dealing with morphological knowledge to respond to vocabulary tasks (Lex30; and G_Lex) for English language learners should be prudent, and learners' ability to distinguish different levels of morphological knowledge relates to their different language proficiency levels. A higher morphological knowledge level (e.g., level 6) will overestimate language learners' vocabulary knowledge, whereas a lower morphology level (e.g., level 1) will underestimate English language learners' vocabulary knowledge. The former can cause low vocabulary scores for English language learners, and the latter can cause high vocabulary scores for language learners. This also applies to LD measures, as overestimating morphological knowledge levels of language learners can reduce LD scores within their writing or vice versa. Therefore, the current chapter uses the flemma (inflected suffixes without distinguishing part-of-speech) for vocabulary tasks (Lex30 and G_Lex) and all writing samples.

4.2.1 Measures

The study reported in the current chapter uses the same three productive vocabulary tasks (Lex30, G_Lex, and the PVLT), as introduced in chapter 3 (see section 3.2.1.1). The current chapter also uses the same multiple lexical diversity measures (Types, D (vocd), HD-D, TTR, Log_TTR, Root_TTR, MSTTR, MAAS, MATTR, MTLD, MTLD_W) as introduced in chapter 3 (see section 3.2.1.2).

4.2.2 Participants

The participants were 91 English as a Foreign Language (EFL) learners from the same language background (L1 Japanese speakers). The L1 Japanese participants were undergraduates from different majors. The proficiency levels for the Japanese participants were B1, B2 and C1 level English learners, as judged by their English language instructors. I asked all participants to respond to the three productive vocabulary tasks and one IELTS writing topic. The participants in the current gave their consent to take the study, and the process followed the ethical procedures. All participants joined the experiment voluntarily, and they reserved the right to withdraw at any time. The Research Ethics Committee of the Graduate School of Humanities and Social Sciences, Hiroshima University has approved this research (approval number: HR-HUM-000804).

The current study used a priori power analysis (G*Power, Faul et al., 2007) to estimate sample size using a two-tailed test. To determine the minimum sample size, we used Cohen's (1988) guidelines on the effect size of the correlation coefficient. G*Power results show that to achieve power (1-β err prob) equal to 0.8 (80% to detect a difference) for a medium effect (Correlation ρ H1 = 0.3) at a significant level 0.05 (α err prob = 0.05), the minimum required sample size is 84. To detect a large effect size (Correlation ρ H1= 0.5) for correlation coefficient at the significant level 0.05 (α err prob = 0.05), the minimum sample size is 29, resulting in an actual power of 0.81 (81% to detect the significance). The sample size (n=91) for my current experimental chapter thus meets and exceeds the medium and large effect size requirement.

4.2.3 Methodology

The L1 Japanese participants completed all tests within two weeks. I asked the participants to complete the three productive vocabulary knowledge tasks in the first week and then the IELTS writing task the following week. The participants completed the

vocabulary tasks and the IELTS writing task with pen and paper. In the current chapter, I only chose the middle 200 English words from their writing for the final analysis. The experiment was conducted in October 2019.

4.2.4 Data Analysis

I converted data into electronic format to meet data processing requirements, as in Treffers-Daller et al. (2018). I corrected all spelling mistakes for the three productive vocabulary tasks (Lex30, G_Lex, and the PVLT) and writing samples. I tolerated the errors of the inflected systems as established in the previously published research (e.g., Treffers-Daller et al., 2018; Treffers-Daller, 2013).

4.2.4.1 Vocabulary Tasks Data Analysis. The scoring standards for the PVLT, Lex30 and G_Lex tasks were all the same as each other in the current chapter, but differed from chapter 3. Chapter 3 followed the same lemma criteria as in Meara and Fitzpatrick (2000), whereas the current chapter used flemmas as the word unit for vocabulary tasks. The flemma script processed all responses from the vocabulary tasks before calculating the scores of vocabulary tasks, as explained in chapter 3 (see section 3.2.4.1). Only Lex30 and G_Lex needed to be flemmatised. Because the PVLT task asked the test-takers to complete predetermined words based on word families, no words needed to be flemmatised in the PVLT task. The Lex30 and G_Lex tasks required test-takers to write the words when they saw the cue words or cue sentences, causing the elicited words to display flemma characteristics. I flemmatised the Lex30 and G_Lex responses using the Python script because it was an efficient way to deal with flemma words. Using the flemma criteria, the POS of homographs did not need to be distinguished (e.g., the verb *can* and noun *can*). The verb *can* and the noun *can* were treated as the same word.

Kristopher Kyle (personal communication) suggested this flemmatisation step. He had developed numerous tools for natural language processing (NLP) relating to computational

linguistics (https://kristopherkyle.github.io/professional-webpage/), and these tools have been used in numerous studies (e.g., Kyle et al., 2018; Kyle & Crossley, 2015). The flemma script uses the tokenize() function, and he also recommended the flemma list, an automated flemma list based on all words in the British National Corpus (BNC), from Laurence Anthony's website (https://www.laurenceanthony.net/). Anthony has developed numerous tools (see the previous link for more information) and published papers in the NLP field (e.g., Anthony, 1999, 2013). The corpus_toolkit package developed by Kyle uses Spacy for tagging and parsing the texts. I also put the data cleaning lines to the Python script while flemmatising the text.

The data cleaning process for the responses in the Lex30 and G_Lex was a necessary step before the calculation of the vocabulary task scores because many responses were not recognised by the website or software, resulting in these words being categorised as off-list, thus resulting in the incorrect calculation of vocabulary task scores. I converted all letters into lowercase. The abbreviations, such as 'TV', 'UN', 'IQ', 'GPS', and 'APP', were written in full spelling. The expressions of numbers written numerically, such as '*1990*', '*50*', '*600*', and so forth, were deleted from the original texts, while the figures expressed as '*fifty*' and '*six hundred*' were kept in the final analysis of the texts. Also, I deleted the names of people, countries, cities, trademarks, foods, months, and weekday expressions from the responses because the knowledge of these words could not be a determining factor in one's vocabulary knowledge.

I followed the same data-cleaning process as described in chapter 3 for processing the three productive vocabulary tasks (the PVLT, Lex30, and G_Lex) since using the same word lists (the BNC/COCA word lists) would ensure the same standard for computing vocabulary task scores. I computed all flemmatised responses from Lex30 and G_Lex and all the correct responses from the PVLT through the AntWordProfiler software created by Anthony (2022)

from this website: https://www.laurenceanthony.net/software/antwordprofiler/. AntWordProfiler is a freeware tool for profiling texts' vocabulary levels and complexity. In the current chapter, I only use AntWordProfiler to score the vocabulary levels by importing the BNC/COCA word family lists created by Nation (2017). Many studies have used and validated the BNC/COCA word lists (e.g., Dang, 2020; Stoeckel & McLean, 2022; Webb et al., 2017). First, the creators derive the BNC/COCA lists from big data, which include written and spoken corpora, and they constantly update the lists. Second, the BNC/COCA lists are based on the frequency of 34 baseword lists in the AntWordProfiler. A baseword1 means the first 1K words, baseword2 means 2K words, and in the current study, I have taken only the words beyond the first 1K band as participants' word knowledge. Third, the BNC/COCA word lists also include the word families of the Academic Word List (AWL) built by Coxhead (2000). Fourth, when using the AntWordProfiler to deal with the responses, the treatment of counting units is by using flemma, and it cannot distinguish homonyms so far (Nation, 2016, p. 135). It meets the requirement of word unit counting for the current study.

After removing the types out of the 1K level in the vocabulary tasks, I computed the percentage scores for the three productive vocabulary tasks. As for Lex30 and G_Lex's elicited responses, they are of different frequencies, including both 1K and non-1K words. The PVLT was created from 2K word families. However, several words were found to belong to 1K when it was processed by the word lists based on the BNC-COCA data. I removed those 1K words during the data processing of the current study.

4.2.4.2 Writing Samples Data Analysis. I conducted the same flemmatisation process on the writing texts as the vocabulary tasks. I treated all writing samples using the same flemma script. First, I corrected the spelling mistakes before using the software to calculate the lexical diversity (LD) measures. This was because the software treated the misspelt words as different tokens, which resulted in inaccurate counting of word types and increased the

scores of the lexical diversity measures. Second, I counted the words from the writing samples like ‘don’t’, ‘doesn’t’, ‘there’re’, ‘it’s’, ‘we’ve’, ‘i’m’, ‘shouldn’t’, ‘can’t’, and ‘isn’t’ as one type. Third, I conducted the data-cleaning process and the flemmatising process simultaneously. I removed some words during the data-cleaning process. Based on Treffers-Daller et al. (2018), to ensure the word length meets 200 English words, it is better to select a relatively longer text (e.g., the middle 220 words) before the data-cleaning phase or to flemmatise the whole text before selecting the number of words to analyse. As in Treffers-Daller et al. (2018), we analysed only the middle 200 words because language learners may paraphrase the writing topic or use formulaic language at the beginning and end of the writing.

After the writing samples had been flemmatised, the next step was to calculate the LD measures’ scores. I computed LD by CLAN (MacWhinney, 2000) for the D (vocd) measure and by Python script for the remaining measures (Types, TTR, Root_TTR, Log_TTR, MSTTR, MAAS, MTLD, MTLD_W, and MATTR). To meet the flemma criteria for the LD measures in the current study, I used the Python script for all LD measures except D (vocd). The writing samples would be coded into two formats: the CHAT format for CLAN and the .txt (UTF-8) format for Python script. I calculated the D (vocd) scores in the mor tier, the morphological analysis designed within the CLAN software. I used the Python script to compute the remaining LD measures.

4.3 Results

The following tables show the results of the productive vocabulary knowledge tasks and LD measures and the relationships between the productive vocabulary knowledge tasks and IELTS writing proficiency.

Table 4.1 shows the descriptive statistics of the three productive vocabulary knowledge tasks: Lex30, G_Lex, and the PVLT. The results in Table 4.1 show that the mean

percentage scores of the three productive vocabulary knowledge tasks differ from each other. Lex30 elicits the highest mean scores (mean=22.84%), followed by the PVLT (mean=21.86%), and then G_Lex (mean=13.31%). Meanwhile, the mean scores between Lex30 and the PVLT are very similar.

Table 4.1

Productive Vocabulary Knowledge (PVK) Tasks Descriptive Statistics

PVK measures (n=91)	Minimum	Maximum	Mean	SD
Lex30%	6.67	61.67	22.84	10.97
G_Lex%	2.50	36.67	13.32	7.97
PVLT%	6.67	72.22	21.86	15.5

The Shapiro-Wilk test results show that the following variables violate the normal distribution assumption ($p<0.05$): Lex30% ($p=0.000$); G_Lex% ($p=0.000$); the PVLT% ($p=0.000$); Log_TTR ($p=0.002$); MAAS ($p=0.000$); MTLD ($p=0.01$); and MTLD-W ($p=0.008$). I ran the nonparametric correlations, the Spearman's rho, for the data which violate the normal distribution and the robust regression analyses using the bootstrapping method.

Table 4.2 shows the correlations between the three productive vocabulary knowledge tasks (Lex30, G_Lex, and the PVLT). The results in Table 4.2 show that the strongest correlations were between G_Lex and the PVLT ($r=.671^{**}$, $p<0.01$), followed by the correlations between Lex30 and the PVLT ($r=.592^{**}$, $p<0.01$), and then the correlations between Lex30 and G_Lex ($r=.590^{**}$, $p<0.01$).

Table 4.2

Correlations Between Productive Vocabulary Knowledge (PVK) Tasks

PVK measures (n=91)	G_Lex%	PVLT%
Lex30%	.590**	.592**
G_Lex%		.671**

Note. **. Significant at the 0.01 level (2-tailed)

Table 4.3 shows the descriptive statistics of the lexical diversity measures. Here, I report the LD mean scores and standard deviations as the following in Table 4.3: the largest mean score is Types (mean=95.68), followed by D (vocd) (mean=48.08), and afterwards MTLD_W (mean=47.28). The highest SD is D (vocd) (SD=13.94), followed by MTLD (12.63), and then MTLD_W (SD=12.51).

Table 4.3

Descriptive Statistics of Lexical Diversity (LD) Measures

LD measures	Minimum	Maximum	Mean	SD
Types	68	118	95.68	12.20
TTR	.34	.60	.48	.06
Root_TTR	4.8	8.41	6.76	.87
Log_TTR	.80	.90	.86	.03
MSTTR	.59	.81	.71	.05
MAAS	.04	.09	.06	.01
D (*vocd*)	23.16	83.87	48.08	13.94
HD-D	.64	.83	.75	.04
MTLD	25.36	73.72	47.13	12.63
MTLD_W	24.96	81.10	47.28	12.51
MATTR	.59	.82	.71	.05

Table 4.4 shows the correlations between lexical diversity (LD) measures. As shown in Table 4.4, LD measures strongly and significantly correlate. The strongest correlations are between Types and Root_TTR (r=1.000**, p<0.01), followed by the correlations between TTR and Root_TTR (r=.999**, p<0.01).

Table 4.4

Correlations Between Lexical Diversity (LD) Measures

LD measures	TTR	Root_TTR	Log_TTR	MSTTR	MAAS	D (vocd)	HD-D	MTLD	MTLD_W	MATTR
Types	.998**	1.000**	.991**	.818**	-.959**	.890**	.901**	.810**	.832**	.838**
TTR		.999**	.990**	.815**	-.958**	.882**	.893**	.808**	.828**	.834**
Root_TTR			.991**	.818**	-.959**	.889**	.900**	.811**	.833**	.838**
Log_TTR				.802**	-.953**	.884**	.886**	.804**	.829**	.826**
MSTTR					-.779**	.834**	.874**	.911**	.933**	.915**
MAAS						-.865**	-.872**	-.783**	-.790**	-.789**
D (vocd)							.968**	.880**	.906**	.853**
HD-D								.882**	.904**	.888**
MTLD									.940**	.889**
MTLD_W										.937**

Note. **. Significant at the 0.01 level (2-tailed)

Table 4.5 shows the correlations between the three productive vocabulary tasks (Lex30, G_Lex, and the PVLT) and lexical diversity scores. Table 4.5 shows significant correlations between all three productive vocabulary tasks and LD measures. The higher the productive vocabulary knowledge task scores, the greater the lexical diversity scores in IELTS writing, and the opposite is also true. The strongest correlations are between the PVLT and MAAS (r=.518**, p<0.01), followed by the correlations between G_Lex and MAAS (r=.510**, p<0.01). In addition, G_Lex and the PVLT show closer relationships with LD measures than Lex30 does, aside from four LD results (Root_TTR; MATTR; MTLD; MTLD_W). There is a slightly enhanced correlation between the PVLT and LD results compared to G_Lex and LD results.

Table 4.5

Correlations Between Productive Vocabulary Tasks Scores and Lexical Diversity (LD) Scores

LD measures	Lex30%	G_Lex%	PVLT%
Types	.345**	.479**	.483**
TTR	.357**	.493**	.503**
Root_TTR	.487**	.487**	.490**
Log_TTR	.356**	.482**	.504**
MSTTR	.271**	.308**	.355**
MAAS	-.362**	-.510**	-.518**
D (*vocd*)	.303**	.359**	.367**
HD-D	.282**	.356**	.358**
MTLD	.313**	.262*	.346**
MTLD_W	.264*	.255*	.347**
MATTR	.289**	.278**	.347**

Note. **. Significant at the 0.01 level (2-tailed). *. Significant at the 0.05 level (2-tailed).

To determine whether the vocabulary knowledge tasks can predict LD scores, I ran a robust, simple standard linear regression analysis using bootstrapping (Larson-Hall, 2015). The results in Table 4.6 show that each of the three vocabulary knowledge tasks can explain variance in lexical diversity scores. The results show G_Lex can explain 35.7% of the variance in TTR scores and 34.2% of the variance in Root_TTR scores. The PVLT can explain 33.6% of the variance in TTR scores. G_Lex and the PVLT can explain a bigger percentage of the variance in LD scores than Lex30.

Table 4.6

Regression Analyses Between Vocabulary Tasks Scores and Lexical Diversity (LD) Scores

variable	R^2	sr^2	Intercept	B	95% Confidence Interval	
					Lower Bound	Upper Bound
Lex30→Type	0.149	0.386	85.882	0.429	0.213	0.645
Lex30→TTR	0.160	0.400	0.425	0.002	0.001	0.003
Lex30→Root_TTR	0.152	0.390	6.051	0.031	0.016	0.046
Lex30→Log_TTR	0.143	0.378	0.839	0.001	0.000	0.001
Lex30→MAAS	0.151	-0.389	0.071	0.000	-0.001	0.000
Lex30→MSTTR	0.077	0.278	0.683	0.001	0.000	0.002
Lex30→D (vocd)	0.141	0.375	37.199	0.477	0.229	0.725
Lex30→HD-D	0.091	0.302	0.725	0.001	0.000	0.002
Lex30→MTLD	0.116	0.341	38.174	0.392	0.164	0.620
Lex30→MTLD_W	0.101	0.317	39.012	0.362	0.134	0.590
Lex30→MATTR	0.088	0.297	0.678	0.001	0.000	0.002
G_Lex→Type	0.334	0.578	83.891	0.885	0.622	1.148
G_Lex→TTR	0.357	0.598	0.415	0.005	0.003	0.006
G_Lex→Root_TTR	0.342	0.585	5.907	0.064	0.045	0.082
G_Lex→Log_TTR	0.317	0.563	0.835	0.002	0.001	0.002
G_Lex→MAAS	0.324	-0.569	0.072	-0.001	-0.001	-0.001
G_Lex→MSTTR	0.142	0.376	0.680	0.002	0.001	0.003
G_Lex→D (vocd)	0.237	0.487	36.738	0.852	0.530	1.173
G_Lex→HD-D	0.170	0.413	0.722	0.002	0.001	0.003
G_Lex→MTLD	0.122	0.350	39.754	0.554	0.241	0.866
G_Lex→MTLD_W	0.168	0.410	38.708	0.643	0.342	0.945
G_Lex→MATTR	0.129	0.359	0.679	0.002	0.001	0.003
PVLT→Type	0.309	0.556	86.117	0.438	0.300	0.575
PVLT→TTR	0.336	0.579	0.426	0.002	0.002	0.003
PVLT→Root_TTR	0.316	0.562	6.068	0.031	0.022	0.041
PVLT→Log_TTR	0.291	0.540	0.840	0.001	0.001	0.001
PVLT→MAAS	0.294	-0.542	0.070	0.000	-0.001	0.000
PVLT→MSTTR	0.149	0.386	0.684	0.001	0.001	0.002
PVLT→D (vocd)	0.190	0.436	39.509	0.392	0.222	0.563
PVLT→HD-D	0.144	0.379	0.729	0.001	0.001	0.002
PVLT→MTLD	0.142	0.377	40.427	0.307	0.148	0.466
PVLT→MTLD_W	0.176	0.419	39.881	0.339	0.184	0.493
PVLT→MATTR	0.131	0.362	0.684	0.001	0.001	0.002

Note. Bootstrap results are based on 2000 bootstrap samples.

4.4 Discussion

The introduction section of the present chapter highlighted the need to include enough participants from higher proficiency levels than those reported in chapter 3. I used three productive vocabulary knowledge tasks (Lex30; G_Lex; and the PVLT) and multiple LD measures for the current study with 91 participants whose proficiency levels ranged from CEFR B1 to C1 (compared to the A2 level of the 29 participants in chapter 3). Since different word unit counts influence different vocabulary task scores and LD scores, I used the same flemmatising process for the responses to both the vocabulary tasks and writing samples to keep the word unit consistent. The research question for the current chapter is: *To what extent does a battery of productive vocabulary tasks predict IELTS writing ability for participants at levels B1 to C1?* The results in the current chapter suggest that the three productive vocabulary knowledge tasks can indeed predict IELTS writing scores to varying degrees. The findings in the current chapter address three points: the significant correlations between three productive vocabulary knowledge tasks and LD measures compared to Treffers-Daller et al. (2018), the different significant correlations and regressions between the three productive vocabulary knowledge tasks, and the strongly significant correlations between LD measures.

First, the results in Table 4.5 and Table 4.6 show that all three productive vocabulary knowledge tasks significantly correlate with the LD scores, showing that productive vocabulary tasks (Lex30, G_Lex, and the PVLT) can, to some extent, predict IELTS writing scores. This finding contrasts with the correlations between productive vocabulary knowledge tasks and LD measures found in Treffers-Daller et al.'s (2018) study. The following Table 4.7 represents a comparison between the current study and Treffers-Daller et al.'s (2018) correlations. The current study incorporates five additional LD measures (Log_TTR; MSTTR; MAAS; MATTR; and MTLD_W) not utilised in Treffers-Daller et al. (2018). The correlation values in the current study exhibit closer relationships than those

presented in Treffers-Daller et al. (2018), especially the performance of G_Lex and the PVLT task in Table 4.6. These findings tentatively imply that G_Lex and the PVLT tasks might better predict writing levels when compared to the vocabulary scores in Treffers-Daller et al. (2018).

Table 4.7

Comparison Between the Current Study and Treffers-Daller et al. (2018)

LD measures	Lex30%	G_Lex%	PVLT%	Vocab scores in Treffers-Daller et al.
Types	.345**	.479**	.483**	.468**
TTR	.357**	.493**	.503**	.470**
Root_TTR	.487**	.487**	.490**	.472**
Log_TTR	.356**	.482**	.504**	
MSTTR	.271**	.308**	.355**	
MAAS	-.362**	-.510**	-.518**	
D (*vocd*)	.303**	.359**	.367**	.319**
HD-D	.282**	.356**	.358**	.309**
MTLD	.313**	.262*	.346**	.331**
MTLD_W	.264*	.255*	.347**	
MATTR	.289**	.278**	.347**	

Second, the correlations between productive vocabulary knowledge tasks in the current study are strongly and significantly correlated. The results in Table 4.2 show the significant correlations between Lex30 and G_Lex (r=.590**, p<0.01); Lex30 and the PVLT (r=.592**, p<0.01); and between G_Lex and the PVLT (r=.671**, p<0.01). The strongest and most significant correlations are between G_Lex and the PVLT (r=.671**, p<0.01),

showing that they might elicit the same quality of English words from the participants. The probable explanation is that both G_Lex and PVLT tasks offered a sentence context to the participants, which can elicit a similar number of English words in quantity and quality, as is supported by previous studies (e.g., Clenton, 2010; Edmonds et al., 2022; Fitzpatrick & Clenton, 2017). I will return to this topic in more detail in the discussion chapter. Briefly, though, G_Lex requires participants to write the first four words in the gaps through the cue sentences, and the words used for the cue sentences are usually highly frequent responses (see Appendix B). Similarly, the PVLT can elicit both quality and quantity of vocabulary knowledge because it asks participants to complete the predetermined word based on frequency for each sentence by giving the first few letters (see Appendix B).

Third, all LD measures are strongly and significantly correlated. The significant correlations between types and Root_TTR represent a perfect correlation (r=1.000**, p<0.01). The results in Table 4.4 support the notion of Jarvis and Hashimoto (2021), who mentioned that the current lexical diversity measures develop from the same construct, the type-variation concept. The strong and significant correlations between LD measures show they appear to be assessing the same construct. I will return to this issue pertaining to the LD construct in my discussion chapter. This further confirms the viewpoints of previous studies (Jarvis, 2013; Jarvis & Hashimoto, 2021, Kyle et al., 2021) about the multi-variate nature of the LD measures. These studies have suggested that the LD measures developed so far are based on a modification between types and tokens and cannot express all features of lexical diversity.

4.4.1 Limitations

A limitation of the current study relates to the length of the text written by the Japanese participants. The essay length of Japanese participants for the current chapter is between 200 and 350 English words. The current chapter analysed LD scores based on the

middle 200 English words, as suggested by Treffers-Daller et al. Despite extracting the middle 200 English words, the LD scores inevitably varied for every 200 words of an essay over 200. To explain, since adjusting the number of types and tokens can determine LD scores, any difference in the proportion of types used for different parts of an essay can cause differences in the LD score, so a random 200-word sample selection from an essay will not reflect an equivalent proportion of types from a specific participant. As a result, ensuring an equal proportion of types of 200-word samples during the writing process is probably impossible, such as if I divide a 2000-word essay into ten sections (e.g., 200 words each), the same types will not be used throughout each section.

4.4.2 Conclusion

The current study has investigated whether productive vocabulary tasks can predict IELTS writing proficiency for L1 Japanese participants (n=91). Their proficiency levels ranged from B1 to C1, as judged by their English language instructors. Thus, it is like Treffers-Daller et al. (2018), where the participants' levels ranged from B1 to C2. The results in the current chapter show significant correlations between the three productive vocabulary tasks (Lex30, G_Lex, and the PVLT) and all lexical diversity measures. Based on the current chapter findings, we can conclude that productive vocabulary tasks, to some extent, predict IELTS writing scores.

I want to distinguish language learners with different IELTS writing levels through three productive vocabulary tasks (Lex30; G_Lex; and the PVLT) in chapter 5 and then investigate whether lexical diversity scores will improve along with the productive vocabulary knowledge scores when I focus on improving participants' vocabulary knowledge in chapter 6. The current chapter divided the levels of participants based on their CEFR level, as judged by their English language instructors. Since the CEFR assesses four different language skills, listening, speaking, reading, and writing, the current study only contains

participants' IELTS writing scores. In the following chapters, I will first examine the IELTS scores in chapter 5 and report on a study in which qualified IELTS raters mark IELTS writing samples. Then, according to the raters' scores, I divide the participants into different proficiency groups. Therefore, the experiment reported in chapter 5 explores whether productive vocabulary tasks can differentiate between different IELTS writing levels. Further, I investigate whether productive vocabulary knowledge task scores and writing scores can change over a short study period in chapter 6 by only focusing on improving participants' vocabulary knowledge.

Chapter 5: To What Extent Can Productive Vocabulary Tasks Differentiate Between IELTS Writing Scores?

5.1 Introduction

The results presented in the experiment reported in chapter 4 showed significant correlations between the three productive vocabulary tasks (Lex30, G_Lex, and the PVLT) and LD measures for participants at levels B1 to C1. Chapter 4 investigated 91 L1 Japanese participants to explore potential relationships between three productive vocabulary tasks (Lex30, G_Lex, and the PVLT) and a spectrum of major lexical diversity measures. Chapter 4 continued exploring the research question addressed in chapter 3 regarding the potential relationships between vocabulary tasks and LD measures by focusing on the productive side, with participants' levels ranging from B1 to C1. The findings in the experiment reported in chapter 4 showed that productive vocabulary tasks could, to some extent, predict writing scores, depending on the productive vocabulary knowledge task in question. The results in chapter 4 raised one major issue: the correlations between the three different productive vocabulary tasks (Lex30, G_Lex, and the PVLT) and the LD measures. In contrast to Treffers-Daller et al. (2018), the correlations between the vocabulary scores (G_Lex, and the PVLT) and LD measures in chapter 4 reached a higher level than the correlations reported in Treffers-Daller et al. (2018). The correlations from the experiment reported in chapter 4 showed that these vocabulary tasks more accurately reflected the participant writing levels than those correlations reported in Treffers-Daller et al. (2018).

The previous experimental chapters (chapters 3 and 4) investigated potential relationships between vocabulary knowledge tasks and lexical diversity measures for participants at different CEFR levels. The current chapter investigates specific IELTS writing scores rather than the CEFR levels for all participants. The motivation for this change in investigation relates to the CEFR criteria, which involve a combination of four different

language skills (listening, speaking, reading, and writing), and which I believe might obfuscate results. To investigate this concern, the current experiment in this chapter will focus on participants' IELTS writing scores. The approach I adopt is not unique, since Daller et al. (2013) used trained IELTS raters to rate the IELTS writing samples in their investigation of whether participants' vocabulary knowledge could improve along with a theorised learning curve. The experiment conducted in the current chapter, following Daller et al.'s (2013) study, uses human raters to rate all writing samples.

To differentiate participants' IELTS writing scores, two qualified IELTS raters rated all the writing samples in the experiment in the current chapter. Employing human raters is a crucial method for judging writing proficiency levels, as asserted in previous studies (e.g., Daller et al., 2013; Jarvis, 2013a, 2013b, 2017; Kyle et al., 2020). Daller et al. (2013) investigated a longitudinal study on vocabulary production with 42 participants who wrote 294 essays within a two-year-long teaching period. Their study employed lexical diversity measures and trained IELTS raters to assess participants' vocabulary knowledge. The IELTS raters in their study scored the writing samples from two aspects: the holistic rating according to IELTS writing scoring criteria, and the lexical rating according to vocabulary use for IELTS writing samples. Their structural equation modelling suggested that the lexical diversity measures could not replace the function of human judgement.

Similarly, Jarvis (2013a, 2013b, 2017) discussed the multidimensional nature of the lexical diversity construct relating to seven properties: volume, abundance, variety, evenness, dispersion, specialness, and disparity. He indicated that the lexical diversity measures developed so far mainly focus on the first three features of LD and cannot capture the whole construct of lexical diversity. In Jarvis's (2017) study, twenty human raters rated both LD scores and CEFR writing scores, and the results showed that the relationship between the LD rating and CEFR writing ratings was r=.89. The high correlations showed that human raters

essentially assessed similar aspects concerning LD scores and CEFR ratings. Kyle et al. (2020) explored three features of LD, including abundance (number of types as lemmas), variety (proportion of unique words; they selected four LD measures relatively independent from text length: HD-D; MATTR; MTLD; and MTLD-W), and volume (number of tokens). Their study suggested that abundance could predict the LD score of human raters most, followed by volume and variety. When taken together, these studies (Daller et al., 2013; Jarvis, 2013a, 2013b, 2017; Kyle et al., 2020) suggest a crucial role for human judgement in writing. Based on such research foundations, the current chapter adopts a rating approach with two qualified IELTS raters judging participant writing samples.

According to the IELTS writing band descriptor (https://takeielts.britishcouncil.org/sites/default/files/ielts_task_2_writing_band_descriptors.pdf, see Appendix C), the IELTS raters judged each writing from band 0 to band 9 in four aspects: task response, coherence and cohesion, lexical resource, and grammatical range and accuracy. The IELTS writing scores are the mean scores of these four different rating scales. The raters gave each participant two IELTS writing scores because they each wrote about two topics. Their final IELTS writing scores are the mean scores based on the results from the two IELTS raters.

Based on the IELTS raters' scores in the writing samples, I divided the participants into three different IELTS writing groups. The current chapter explores whether productive vocabulary tasks (Lex30, G_Lex, and the PVLT) can differentiate between IELTS writing scores. The results in the experiment in chapter 4 showed that the correlations between productive vocabulary tasks and lexical diversity measures varied in strength, ranging from weak to moderate correlations: Lex30<G_Lex<the PVLT. Likewise, the three productive vocabulary tasks are also engaged contextually in this sequence: Lex30<G_Lex<the PVLT. The Lex30 task offers the spelling context, the G_Lex task provides semantic context, and the

PVLT requires semantic, collocational, and syntactic knowledge. The PVLT task involves the most context when compared to G_Lex and Lex30 (see Table 5.1 for more information about the contextual engagement of the three vocabulary knowledge tasks).

Table 5.1

Dimensions of Vocabulary Knowledge Tapped by Three Vocabulary Tests

				Lex30	G_Lex	PVLT
Form	spoken	R	What does the word sound like?			
		P	How is the word pronounced?			
	written	R	What does the word look like?			✓
		P	How is the word written and spelled?	✓	✓	✓
	word parts	R	What parts are recognizable in this word?			
		P	What word parts are needed to express the meaning?			
Meaning	form and meaning	R	What meaning does this word form signal?			
		P	What word form can be used to express this meaning?	✓	✓	✓
	concept and referents	R	What is included in the concept?			
		P	What items can the concept refer to?			
	associations	R	What other words does this make us think of?			
		P	What other words could we use instead of this one?	✓	✓	
Use	grammatical functions	R	In what patterns does the word occur?			✓
		P	In what patterns must we use this word?			✓
	collocations	R	What words or types of words occur with this one?		✓	✓
		P	What words or types of words must we use with this word?			
	constraints on use	R	Where, when, and how often would we expect to meet this			
		P	Where, when, and how often can we use this word?			✓

Note. R = receptive knowledge, P = productive knowledge. Reprinted from "Exploring the construct validity of tests used to assess L2 productive vocabulary knowledge," by A. Edmonds, J. Clenton, and H. Elmetaher, 2022, *System*, *108*, 102855, p. 4 (https://doi.org/10.1016/j.system.2022.102855). Copyright 2022 by the Elsevier Ltd.

Assuming higher-level participants possess greater vocabulary knowledge than lower-level participants do, participants with more vocabulary knowledge should have greater language ability to build lexical networks. I hypothesise that, for higher-level participants, closer relationships between the three productive vocabulary tasks and LD measures can be expected than for lower-level participants. According to Milton (2013), 'once a meaning is attached to that form and some idea is gained as to how the word can be used, then it develops links with other words and begins to network and it does not matter whether these are grammatical or associational or collocational links' (p. 61).

Thus, the research questions set for the current chapter are:

RQ1: *To what extent can productive vocabulary tasks differentiate between IELTS writing scores for participants at levels B1 to C1?*

RQ2: *Do the results from a comparison of productive vocabulary tasks and lexical diversity measures reflect an increase in writing scores?*

5.2 Study

The current chapter examines whether productive vocabulary tasks can distinguish between different IELTS writing scores among participants. Qualified IELTS raters marked all the writing samples based on the IELTS writing rubric (see Appendix C). I divided all participants from the two different language backgrounds into three proficiency groups based on their IELTS writing scores. The experiment reported in the current chapter uses the same word unit (flemma) as chapter 4 for productive vocabulary tasks and IELTS writing samples. The flemma standards and flemma lists are the same as described in chapter 4 (see section 4.2.4).

5.2.1 Measures

The current chapter uses the same three productive vocabulary tasks (Lex30, G_Lex, and the PVLT) introduced in section 3.2.1.1. The current chapter also uses the same spectrum

of lexical diversity measures, Types, TTR, Root_TTR, Log_TTR, MSTTR, MAAS, D (vocd), HD-D, MTLD, MTLD_W, MATTR, as reported in chapter 3 and chapter 4. I presented the introduction to lexical diversity measures in section 3.3.1.2.

5.2.2 Participants

The participants were 98 English language learners from two different language backgrounds (63 L1 Japanese speakers and 35 L1 French speakers). The L1 Japanese participants were undergraduates from three different majors. I collected the L1 Japanese data in Japan. Because COVID was limiting physical access to participants and because of the need for higher-level participants for the current chapter, I sought help from a colleague at a French university to collect data. Their instructor distributed a paper copy of all tasks to L1 French (L2 English learner) participants, and the colleague scanned all completed responses and returned them to me by email. I then typed all the responses into an electronic format. The L1 French participants were undergraduates in an English program course. The participants were aged between 18 to 20 years old. The instructor asked participants to finish the three productive vocabulary tasks, Lex30, G_Lex, and the PVLT, as well as two IELTS writing topics. The IELTS writing levels for all 98 participants ranged from the intermediate (B1/B2) to advanced (C1) levels. After I collected the data, following Daller et al.'s (2013) study, qualified IELTS raters judged all IELTS samples. According to the IELTS writing results from the raters, I divided the participants into three groups (see Table 5.2), and most Japanese participants belonged to the intermediate level. In contrast, most French participants were at the advanced level.

Table 5.2

IELTS Writing Scores Based on IELTS Ratings

IELTS writing scores	5.5	6	≥6.5
N	28	49	21
N (L1 Japanese)	27	30	6
N (L1 French)	1	19	15

The participants in the current chapter gave their consent to do the study, and the process followed the ethical procedures. All participants took the experiment voluntarily and reserved the right to withdraw at any time (see Appendix D for the ethical convention). The Research Ethics Committee of the Graduate School of Humanities and Social Sciences, Hiroshima University, has approved this research (approval number: HR-HUM-000804).

To determine the minimum sample size, we used Cohen's (1988) guidelines on the effect size of the correlation coefficient. G*Power results show that to achieve power (1-β err prob) equal to 0.8 (80% to detect a difference) for a medium effect (Correlation ρ H1 = 0.3) at a significant level 0.05 (α err prob = 0.05), the minimum required sample size is 84. To detect a large effect size (Correlation ρ H1= 0.5) for correlation coefficient at the significant level 0.05 (α err prob = 0.05), the minimum sample size is 29, resulting in an actual power of 0.81 (81% to detect the significance). The sample size (n=98) for my current experimental chapter meets and exceeds the medium and large effect size requirement.

5.2.3 Methodology

I asked all participants to complete the productive vocabulary knowledge tasks (Lex30, G_Lex, and the PVLT) and two IELTS writing topics within two weeks (see Appendix E). I used pen and paper for data collection, and their instructors controlled the testing time. I gave all participants the paper format for the three productive vocabulary

knowledge tasks and two IELTS writing topics. They completed all tasks during class time. The experiment was conducted in April 2020.

5.2.4 Data Analysis

The data analysis procedures for the current chapter were the same as described in chapter 4 (see section 4.2.4 for detailed information on data analysis). I used the same data processing procedures for the three productive vocabulary tasks (Lex30, G_Lex, and the PVLT), as reported in section 4.2.4.1. Meanwhile, I used the same procedures to process the IELTS writing samples in section 4.2.4.2. The only difference between the current chapter and chapter 4 is the number of IELTS writing samples. Chapter 4 collected one IELTS writing sample from each participant, whereas the current chapter used two. I calculated the mean scores of the lexical diversity measures across two different IELTS writing topics. In the current chapter, I only chose the middle 200 English words for the final analysis.

5.3 Results

The following tables show the results for the three productive vocabulary knowledge tasks and LD measures and the relationships between productive vocabulary knowledge tasks and LD measures. In addition, I report on whether three productive vocabulary knowledge tasks can distinguish IELTS writing scores by exploring the correlations between the productive vocabulary knowledge tasks and LD measures at different IELTS writing scores.

Table 5.3 shows the descriptive statistics of the three productive vocabulary tasks (Lex30, G_Lex, and the PVLT). As shown in Table 5.3, the mean percentage scores differ between the three productive vocabulary knowledge tasks. The PVLT elicited the largest mean percentage scores (mean%=27.24), and then Lex30 (mean%=22.84), and this is followed by G_Lex (mean%=11.32).

Table 5.3

Productive Vocabulary Knowledge (PVK) Tasks Descriptive Statistics

PVK measures (n=98)	Minimum	Maximum	Mean	SD
Lex30%	8.33	50.00	22.84	8.91
G_Lex%	2.50	30.00	11.32	5.71
PVLT%	7.78	75.56	27.24	19.18

The Shapiro-Wilk tests show that some variables violate the normal distribution ($p<0.05$), and I ran the non-parametric analysis (Spearman's rho) for these variables and robust regression analyses using bootstrapping.

Table 5.4 shows the correlations between the three productive vocabulary knowledge tasks (Lex30, G_Lex, and the PVLT). The results in Table 5.4 show that the strongest significant correlations were between Lex30 and the PVLT (r_s=.689**, $p<0.01$), followed by the significant correlations between Lex30 and G_Lex (r_s=.581**, $p<0.01$), and then the significant correlations between G_Lex and the PVLT (r_s=.476**, $p<0.01$).

Table 5.4

Correlations Between Productive Vocabulary Knowledge (PVK) Tasks

PVK measures (n=98)	G_Lex%	PVLT%
Lex30%	.581**	.689**
G_Lex%		.476**

Note. **. Significant at the 0.01 level (2-tailed)

Table 5.5 shows the descriptive statistics of lexical diversity measures. The results in Table 5.5 show the mean scores and standard deviations of lexical diversity measurements for all participants (n=98). The highest mean score is Types (mean=97.81), followed by MTLD

(mean=53.70), and then MTLD-W (mean=53.16). The highest SD is D (vocd) (SD=13.90), followed by MTLD (SD=13.85), and afterwards MTLD-W (SD=13.50).

Table 5.5

Descriptive Statistics of Lexical Diversity (LD) Measures

LD measures (n=98)	Minimum	Maximum	Mean	SD
Types	67	119	97.81	11.21
TTR	.33	.59	0.49	0.06
Root_TTR	4.69	8.39	7.00	.79
Log_TTR	.79	.90	.86	.02
MSTTR	.58	.81	.73	.05
MAAS	.04	.09	0.06	0.01
D (*vocd*)	21.15	85.94	52.85	13.90
HD-D	.62	.84	0.76	0.04
MTLD	28.56	94.20	53.70	13.85
MTLD-W	28.30	92.34	53.16	13.50
MATTR	.58	.82	0.73	0.05

Table 5.6 shows the correlations between lexical diversity measures for the 98 participants. These correlations are strongly and significantly correlated, and the correlations between some LD measures reached absolute positive correlations. The absolute correlations were between Types, TTR, Root_TTR, Log_TTR, and MAAS (r=1.000**/-1.000**, $p<0.01$).

Table 5.6

Correlations Between Lexical Diversity (LD) Measures

LD measures (n=98)	TTR	Root_TTR	Log_TTR	MSTTR	MAAS	D (*vocd*)	HD-D	MTLD	MTLD-W	MATTR
Types	1.000**	1.000**	1.000**	.914**	-1.000**	.940**	.941**	.900**	.906**	.917**
TTR		1.000**	1.000**	.914**	-1.000**	.940**	.941**	.900**	.906**	.917**
Root_TTR			1.000**	.914**	-1.000**	.940**	.941**	.901**	.906**	.917**
Log_TTR				.888**	-1.000**	.941**	.942**	.901**	.907**	.895**
MSTTR					-.888**	.927**	.924**	.964**	.975**	.972**
MAAS						-.941**	-.942**	-.901**	-.907**	-.895**
D (*vocd*)							.995**	.914**	.934**	.921**
HD-D								.910**	.929**	.914**
MTLD									.970**	.960**
MTLD-W										.970**

Note. **. Significant at the 0.01 level (2-tailed)

Table 5.7 shows the correlations between the percentage scores of the three productive vocabulary tasks (Lex30, G_Lex, and the PVLT) and lexical diversity measures. Table 5.7 shows that there were correlations between the vocabulary tasks and measures of lexical diversity. The correlation results show the positive relationships between the three productive vocabulary tasks and lexical diversity measures. The strongest correlations were between three LD measures (Types, TTR, and Root_TTR) and the PVLT (r=.693**, p<0.01), followed by the correlations between two LD measures (Log_TTR and MAAS) and the PVLT (r=.690**/-.690**, p<0.01). The PVLT task has the strongest correlations with all LD measures, followed by Lex30 and LD measures, and afterwards, G_Lex and LD measures.

Table 5.7

Correlations Between Productive Vocabulary Tasks Scores and Lexical Diversity (LD) Scores

LD measures (n=98)	Lex30%	G_Lex%	PVLT%
Types	.603**	.403**	.693**
TTR	.603**	.403**	.693**
Root_TTR	.603**	.404**	.693**
Log_TTR	.602**	.403**	.690**
MSTTR	.576**	.332**	.636**
MAAS	-.602**	-.403**	-.690**
D (*vocd*)	.557**	.354**	.661**
HD-D	.544**	.353**	.649**
MTLD	.596**	.325**	.612**
MTLD-W	.578**	.305**	.633**
MATTR	.568**	.300**	.620**

Note. **. Significant at the 0.01 level (2-tailed)

Table 5.8 shows participants' mean lexical diversity scores at different IELTS writing levels. According to their IELTS writing scores, I divided the participants into three groups: those whose IELTS writing scores are 5.5, 6.0, and those whose IELTS writing scores are equal to or over 6.5. As shown in Table 5.8, participants with higher lexical diversity scores achieved higher proficiency levels in their IELTS writing. For example, participants whose IELTS writing scores were equal to or over 6.5 have a higher LD score than those whose IELTS writing scores were 6.0 or 5.5. Moreover, participants whose IELTS writing scores are 6.0 had a higher LD score than those with an IELTS writing score of 5.5.

Table 5.8

Lexical Diversity Measures Scores at Different IELTS Writing Scores

Measures	IELTS writing = 5.5 (n=28)	IELTS writing = 6.0 (n=49)	IELTS writing ≥6.5 (n=21)
Types	92.77	96.76	106.98
TTR	.46	.48	.53
Root_TTR	6.54	6.83	7.55
Log_TTR	.85	.86	.88
MSTTR	.71	.72	.76
MAAS	.06	.06	.05
D (*vocd*)	45.97	51.78	64.51
HD-D	.75	.76	.80
MTLD	47.15	52.43	65.42
MTLD-W	46.05	51.84	65.71
MATTR	.71	.73	.76

Table 5.9 shows the correlations between the three productive vocabulary knowledge tasks (Lex30, G_Lex, and the PVLT) and lexical diversity scores for participants with three

different IELTS writing scores: 5.5, 6.0, and equal to or above 6.5. Table 5.9 shows that all vocabulary tasks (Lex30, G_Lex and the PVLT) show significant correlations with IELTS writing scores only for participants whose IELTS scores fall at 6.0 and equal to or over 6.5, but not for those whose IELTS writing scores are 5.5. Second, Table 5.9 shows that the overall correlations between the three productive vocabulary knowledge tasks (Lex30, G_Lex, and the PVLT) and IELTS writing scores increase along with the increasing IELTS writing scores. The strength of significant correlations follows the same sequence as their IELTS writing scores. The strength of correlation between productive vocabulary knowledge scores and lexical diversity scores will increase with the increasing IELTS writing scores. Third, the strengths of the significant correlations differ among Lex30, G_Lex, and the PVLT. The strongest significant correlations were with IELTS writing scores equal to or above 6.5 between MSTTR and the PVLT ($r=.798^{**}$, $p<0.01$), followed by the significant correlations between MATTR and the PVLT ($r=.758^{**}$, $p<0.01$). In addition, the PVLT shows the strongest significant correlations with participants across the three different IELTS writing scores, followed by the Lex30 task and G_Lex. Specifically, the PVLT and Lex30 show significant correlations with IELTS writing scores at 6.0 and equal to or over 6.5, and G_Lex shows moderately significant correlations with IELTS writing scores at 6.0 and equal to or over 6.5. Fourth, for participants whose IELTS writing scores fall at 5.5, there are no significant correlations between vocabulary knowledge scores and LD scores.

Table 5.9

Correlations between Productive Vocabulary Tasks and Lexical Diversity Measures at Different IELTS Writing Scores

Scores	PVK	Types	TTR	Root_TTR	Log_TTR	MSTTR	MAAS	D (vocd)	HD-D	MTLD	MTLD-W	MATTR
5.5 (N=28)	Lex30	.018	.018	.018	.018	.155	-.018	.128	.057	.156	.106	.191
	G_Lex	-.183	-.183	-.183	-.170	-.186	.170	-.156	-.140	-.190	-.222	-.227
	PVLT	.077	.076	.079	.081	.068	-.081	.062	.052	-.012	.041	.018
6.0 (N=49)	Lex30	.579**	.584**	.583**	.580**	.516**	-.580**	.476**	.479**	.575**	.524**	.478**
	G_Lex	.527**	.527**	.527**	.523**	.404**	-.523**	.472**	.474**	.412**	.365**	.404**
	PVLT	.652**	.655**	.654**	.650**	.602**	-.650**	.614**	.602**	.584**	.601**	.588**
≥6.5 (N=21)	Lex30	.741**	.741**	.741**	.748**	.735**	-.748**	.669**	.694**	.658**	.700**	.745**
	G_Lex	.503*	.502*	.503*	.503*	.540*	-.503*	.422	.435*	.451*	.482*	.475*
	PVLT	.659**	.656**	.656**	.663**	.798**	-.663**	.643**	.656**	.692**	.692**	.758**

**. Significant at the 0.01 level (2-tailed). *. Significant at the 0.05 level (2-tailed).

To examine whether the vocabulary knowledge tasks can predict LD scores, I ran robust simple standard linear regression analyses using bootstrapping (Larson-Hall, 2015), as shown in Table 5.10 and Table 5.11.

Bootstrapping was introduced to the second language acquisition (SLA) field by Larson-Hall and Herrington (2010) as a robust statistical analysis method for non-parametric data, or low power data/ data with small sample sizes; they showed, for instance, that bootstrapping even works for tiny sample sizes (n=10). Bootstrapping can re-sample data which violates the normal distribution assumption. LaFlair et al. (2015) offered a guide for different statistics: descriptive statistics, t-tests, ANOVAs, and correlations. Plonsky et al. (2015) investigated bootstrapping by reanalysing 26 published studies from two high impact applied linguistics journals. Their paper found inconsistencies with the original results reported in the papers that offered the raw data. Plonsky et al.'s paper recommended using bootstrapping for data that violated parametric assumptions. However, because the paper offered to them mainly used t-tests and ANOVAs, they examined bootstrapping mainly with t-tests and ANOVAs. Bootstrapping with other statistics, such as correlations and regressions, has not been examined, and they recommended this future empirical studies.

McLean et al. (2020) used a bootstrapping method to investigate the relationships between L2 reading proficiency and vocabulary knowledge. Hamrick (2019) evaluated the overfitting issue for the L2 research using regression analysis and validated this issue through bootstrapping. He set the bootstrapping at 5,000 samples. The results of Hamrick's paper showed that using linear regression analysis can overfit the model for the 'simple linear regression model'. The results of Hamrick's study also highlighted the importance of bootstrapping when doing simple linear regression analysis.

Considering that the data for the current chapter also violated the assumptions for conducting linear regression, I followed these earlier published papers and conducted the linear regression using bootstrapping with a 2000 sample (Larson-Hall, 2015).

Table 5.10 shows the regression analyses between vocabulary tasks and lexical diversity scores for all participants (n=98). The results in Table 5.10 show that vocabulary knowledge scores can predict lexical diversity scores. The R^2 values presented in Table 2 show the extent to which each vocabulary score can account for the variance observed in writing scores. Specifically, the PVLT can explain the largest percentage of variance in lexical diversity scores, followed by Lex30 and G_Lex. The PVLT scores account for the largest proportion of variance in the TTR score (56.9%).

Table 5.10

Regression Analyses Between Vocabulary Tasks Scores and Lexical Diversity (LD) Scores for All Participants (n=98)

Variable (n=98)	R^2	sr^2	Intercept	B	95% Confidence Interval (CI)	
					Lower Bound	Upper Bound
Lex30→Type	0.369	0.607	80.366	0.764	0.561	0.966
Lex30→TTR	0.369	0.608	0.400	0.004	0.003	0.005
Lex30→Root_TTR	0.369	0.607	5.668	0.054	0.040	0.068
Lex30→Log_TTR	0.345	0.587	0.829	0.001	0.001	0.002
Lex30→MAAS	0.345	-0.587	0.074	-0.001	-0.001	0.000
Lex30→MSTTR	0.322	0.568	0.662	0.003	0.002	0.004
Lex30→D (vocd)	0.352	0.593	31.716	0.925	0.671	1.179
Lex30→HD-D	0.281	0.530	0.710	0.002	0.002	0.003
Lex30→MTLD	0.396	0.629	31.377	0.978	0.733	1.222
Lex30→MTLD_W	0.409	0.640	31.027	0.969	0.733	1.205
Lex30→MATTR	0.319	0.564	0.663	0.003	0.002	0.004
G_Lex→Type	0.189	0.434	88.158	0.852	0.494	1.211
G_Lex→TTR	0.189	0.434	0.439	0.004	0.002	0.006
G_Lex→Root_TTR	0.189	0.434	6.218	0.060	0.035	0.085
G_Lex→Log_TTR	0.175	0.419	0.844	0.002	0.001	0.002
G_Lex→MAAS	0.175	-0.419	0.068	-0.001	-0.001	0.000
G_Lex→MSTTR	0.139	0.373	0.694	0.003	0.001	0.004
G_Lex→D (vocd)	0.173	0.416	41.394	1.012	0.563	1.460
G_Lex→HD-D	0.140	0.375	0.735	0.003	0.001	0.004
G_Lex→MTLD	0.174	0.417	42.267	1.011	0.564	1.457
G_Lex→MTLD_W	0.182	0.427	41.744	1.008	0.575	1.442
G_Lex→MATTR	0.123	0.351	0.698	0.003	0.001	0.004
PVLT→Type	0.568	0.754	85.803	0.441	0.363	0.518
PVLT→TTR	0.569	0.754	0.427	0.002	0.002	0.003
PVLT→Root_TTR	0.568	0.754	6.052	0.031	0.026	0.037
PVLT→Log_TTR	0.531	0.728	0.840	0.001	0.001	0.001
PVLT→MAAS	0.531	-0.728	0.070	0.000	0.000	0.000
PVLT→MSTTR	0.442	0.665	0.685	0.002	0.001	0.002
PVLT→D (vocd)	0.534	0.731	38.414	0.530	0.430	0.630
PVLT→HD-D	0.454	0.674	0.726	0.001	0.001	0.002
PVLT→MTLD	0.537	0.733	39.295	0.529	0.429	0.628
PVLT→MTLD_W	0.557	0.746	38.853	0.525	0.430	0.620
PVLT→MATTR	0.431	0.657	0.687	0.002	0.001	0.002

Note. Bootstrap results are based on 2000 bootstrap samples.

Table 5.11 shows the regression analyses between vocabulary task scores and lexical diversity scores for participants of three IELTS writing levels as divided by human raters. The results in Table 5.11 show that vocabulary knowledge scores can predict lexical diversity scores. The R^2 values presented in Table 5.11 indicate the extent to which each vocabulary score can account for the variance observed in writing scores. As writing levels improve, vocabulary scores account for a greater proportion of the variance in lexical diversity scores.

For the participants with an IELTS writing level of 5.5, vocabulary knowledge tasks show a minor percentage of variance in lexical diversity scores. 14.5% of the variance in three lexical diversity measures (Types, TTR, and Root_TTR) can be explained by the PVLT. For the participants with an IELTS writing level of 6.0, the PVLT can explain the largest percentage of variance in lexical diversity scores, followed by Lex30 and G_Lex. The PVLT scores account for an equal proportion of the variance in both TTR and Root_TTR scores, representing 47% of each variance. For the participants with an IELTS writing level of 6.5 or higher, the results show that the three vocabulary knowledge tasks can explain different variances in lexical diversity scores. The PVLT can also explain the highest proportion of variance in lexical diversity scores, followed by Lex30 and G_Lex. The R^2 values show that the PVLT can explain 77.7% of the variance in MSTTR, which is the largest proportion of variance among all vocabulary knowledge scores.

Table 5.11

Regression Analyses Between Vocabulary Tasks Scores and Lexical Diversity (LD) Scores for Participants in Three IELTS Writing Levels

Variable (n=28) IELTS writing level=5.5	R^2	sr^2	Intercept	B	95% Confidence Interval (CI)	
					Lower Bound	Upper Bound
Lex30→Type	0.000	0.018	92.288	0.023	-0.506	0.553
Lex30→TTR	0.000	0.018	0.459	0.000	-0.003	0.003
Lex30→Root_TTR	0.000	0.018	6.510	0.002	-0.036	0.039
Lex30→Log_TTR	0.000	0.018	0.852	0.000	-0.001	0.001
Lex30→MAAS	0.000	-0.018	0.064	0.000	-0.001	0.000
Lex30→MSTTR	0.024	0.155	0.689	0.001	-0.002	0.003
Lex30→D (vocd)	0.016	0.128	41.902	0.199	-0.420	0.818
Lex30→HD-D	0.003	0.057	0.740	0.000	-0.002	0.002
Lex30→MTLD	0.024	0.156	42.166	0.244	-0.377	0.864
Lex30→MTLD_W	0.011	0.106	42.977	0.150	-0.416	0.716
Lex30→MATTR	0.037	0.191	0.684	0.001	-0.001	0.004
G_Lex→Type	0.034	-0.183	96.267	-0.339	-1.072	0.394
G_Lex→TTR	0.034	-0.183	0.479	-0.002	-0.005	0.002
G_Lex→Root_TTR	0.034	-0.183	6.790	-0.024	-0.076	0.028
G_Lex→Log_TTR	0.029	-0.170	0.860	-0.001	-0.002	0.001
G_Lex→MAAS	0.029	0.170	0.061	0.000	0.000	0.001
G_Lex→MSTTR	0.035	-0.186	0.725	-0.002	-0.005	0.002
G_Lex→D (vocd)	0.024	-0.156	49.468	-0.339	-1.206	0.528
G_Lex→HD-D	0.020	-0.140	0.758	-0.001	-0.004	0.002
G_Lex→MTLD	0.036	-0.190	51.462	-0.417	-1.285	0.450
G_Lex→MTLD_W	0.049	-0.222	50.605	-0.441	-1.222	0.340
G_Lex→MATTR	0.051	-0.227	0.730	-0.002	-0.006	0.001
PVLT→Type	0.145	0.381	85.478	0.428	0.009	0.847
PVLT→TTR	0.145	0.381	0.425	0.002	0.000	0.004
PVLT→Root_TTR	0.145	0.381	6.029	0.030	0.001	0.060
PVLT→Log_TTR	0.122	0.349	0.839	0.001	0.000	0.002
PVLT→MAAS	0.122	-0.349	0.070	0.000	-0.001	0.000
PVLT→MSTTR	0.053	0.229	0.688	0.001	-0.001	0.003
PVLT→D (vocd)	0.076	0.277	39.738	0.366	-0.147	0.879
PVLT→HD-D	0.053	0.230	0.729	0.001	-0.001	0.003
PVLT→MTLD	0.055	0.234	41.842	0.312	-0.210	0.834
PVLT→MTLD_W	0.057	0.240	41.129	0.289	-0.183	0.762
PVLT→MATTR	0.029	0.171	0.693	0.001	-0.001	0.003

Note. Bootstrap results are based on 2000 bootstrap samples.

Variable (n=49) IELTS writing level=6.0	R^2	sr^2	Intercept	B	95% Confidence Interval (CI)	
					Lower Bound	Upper Bound
Lex30→Type	0.363	0.602	79.961	0.770	0.471	1.070
Lex30→TTR	0.364	0.603	0.398	0.004	0.002	0.005
Lex30→Root_TTR	0.364	0.603	5.639	0.054	0.033	0.076
Lex30→Log_TTR	0.334	0.578	0.828	0.001	0.001	0.002
Lex30→MAAS	0.333	-0.577	0.075	-0.001	-0.001	0.000
Lex30→MSTTR	0.291	0.539	0.662	0.003	0.002	0.004
Lex30→D (vocd)	0.324	0.569	32.367	0.890	0.513	1.268
Lex30→HD-D	0.258	0.508	0.709	0.002	0.001	0.004
Lex30→MTLD	0.405	0.637	31.933	0.940	0.606	1.274
Lex30→MTLD_W	0.414	0.643	31.933	0.913	0.594	1.232
Lex30→MATTR	0.261	0.511	0.668	0.003	0.001	0.004
G_Lex→Type	0.258	0.508	84.292	1.167	0.586	1.748
G_Lex→TTR	0.259	0.509	0.419	0.006	0.003	0.009
G_Lex→Root_TTR	0.258	0.508	5.945	0.082	0.041	0.123
G_Lex→Log_TTR	0.239	0.488	0.836	0.002	0.001	0.003
G_Lex→MAAS	0.238	-0.488	0.071	-0.001	-0.002	0.000
G_Lex→MSTTR	0.169	0.411	0.682	0.004	0.001	0.006
G_Lex→D (vocd)	0.246	0.496	36.863	1.397	0.680	2.114
G_Lex→HD-D	0.205	0.453	0.720	0.004	0.002	0.006
G_Lex→MTLD	0.227	0.476	38.921	1.265	0.580	1.949
G_Lex→MTLD_W	0.226	0.475	38.889	1.213	0.554	1.871
G_Lex→MATTR	0.166	0.407	0.685	0.004	0.001	0.006
PVLT→Type	0.469	0.685	85.996	0.410	0.282	0.538
PVLT→TTR	0.470	0.686	0.428	0.002	0.001	0.003
PVLT→Root_TTR	0.470	0.685	6.065	0.029	0.020	0.038
PVLT→Log_TTR	0.438	0.662	0.840	0.001	0.001	0.001
PVLT→MAAS	0.438	-0.662	0.070	0.000	0.000	0.000
PVLT→MSTTR	0.331	0.575	0.686	0.001	0.001	0.002
PVLT→D (vocd)	0.425	0.652	39.239	0.478	0.315	0.641
PVLT→HD-D	0.376	0.613	0.726	0.001	0.001	0.002
PVLT→MTLD	0.426	0.653	40.590	0.451	0.297	0.605
PVLT→MTLD_W	0.426	0.653	40.455	0.434	0.286	0.582
PVLT→MATTR	0.320	0.566	0.690	0.001	0.001	0.002

Note. Bootstrap results are based on 2000 bootstrap samples.

Variable (n=21) IELTS writing level ≥6.5	R^2	sr^2	Intercept	B	95% Confidence Interval (CI)	
					Lower Bound	Upper Bound
Lex30→Type	0.549	0.741	86.609	0.717	0.405	1.029
Lex30→TTR	0.548	0.741	0.431	0.004	0.002	0.005
Lex30→Root_TTR	0.549	0.741	6.109	0.051	0.029	0.073
Lex30→Log_TTR	0.559	0.748	0.842	0.001	0.001	0.002
Lex30→MAAS	0.559	-0.748	0.069	-0.001	-0.001	0.000
Lex30→MSTTR	0.540	0.735	0.697	0.002	0.001	0.003
Lex30→D (vocd)	0.447	0.669	41.264	0.818	0.381	1.255
Lex30→HD-D	0.481	0.694	0.740	0.002	0.001	0.003
Lex30→MTLD	0.433	0.658	39.876	0.899	0.405	1.393
Lex30→MTLD_W	0.490	0.700	40.375	0.892	0.455	1.329
Lex30→MATTR	0.555	0.745	0.690	0.003	0.001	0.004
G_Lex→Type	0.253	0.503	97.981	0.637	0.111	1.162
G_Lex→TTR	0.252	0.502	0.487	0.003	0.001	0.006
G_Lex→Root_TTR	0.253	0.503	6.911	0.045	0.008	0.082
G_Lex→Log_TTR	0.253	0.503	0.863	0.001	0.000	0.002
G_Lex→MAAS	0.253	-0.503	0.059	-0.001	-0.001	0.000
G_Lex→MSTTR	0.292	0.540	0.732	0.002	0.001	0.004
G_Lex→D (vocd)	0.178	0.422	54.954	0.676	-0.021	1.373
G_Lex→HD-D	0.189	0.435	0.773	0.002	0.000	0.003
G_Lex→MTLD	0.204	0.451	54.019	0.807	0.041	1.573
G_Lex→MTLD_W	0.232	0.482	54.352	0.804	0.102	1.506
G_Lex→MATTR	0.226	0.475	0.734	0.002	0.000	0.004
PVLT→Type	0.725	0.851	89.642	0.401	0.282	0.520
PVLT→TTR	0.724	0.851	0.446	0.002	0.001	0.003
PVLT→Root_TTR	0.724	0.851	6.323	0.028	0.020	0.037
PVLT→Log_TTR	0.726	0.852	0.848	0.001	0.001	0.001
PVLT→MAAS	0.726	-0.852	0.066	0.000	0.000	0.000
PVLT→MSTTR	0.777	0.882	0.705	0.001	0.001	0.002
PVLT→D (vocd)	0.656	0.810	43.647	0.483	0.315	0.650
PVLT→HD-D	0.686	0.828	0.747	0.001	0.001	0.001
PVLT→MTLD	0.653	0.808	42.184	0.538	0.349	0.726
PVLT→MTLD_W	0.676	0.822	43.656	0.510	0.341	0.680
PVLT→MATTR	0.752	0.867	0.701	0.001	0.001	0.002

Note. Bootstrap results are based on 2000 bootstrap samples.

5.4 Discussion

The purpose of the current chapter has been to investigate whether productive vocabulary tasks can differentiate between IELTS writing scores. I assigned three productive vocabulary tasks (Lex30, G_Lex, and the PVLT) and two IELTS writing topics to all participants (63 L1 Japanese and 35 L1 French), with their IELTS writing scores ranging from B1 to C1. The introduction part of the current chapter highlighted the importance of using qualified IELTS raters to rate participants' IELTS writing. This chapter used two qualified IELTS raters to rate IELTS writing. Based on the raters' results, I divided the participants into three groups according to their IELTS writing scores. The results reported in this chapter show that all three productive vocabulary knowledge tasks can differentiate between IELTS writing scores. The research questions for this chapter were:

RQ1: *To what extent can productive vocabulary tasks differentiate between IELTS writing scores for participants at levels B1 to C1?*

RQ2: *Do the results from a comparison of productive vocabulary tasks and lexical diversity measures reflect an increase in writing scores?*

First, the significant correlations and the regression analyses between the three productive vocabulary knowledge tasks and the LD measures across the three different IELTS writing scores show that the productive vocabulary tasks can, to some extent, differentiate between IELTS writing scores. The results in Table 5.9 show that the significant correlations between productive vocabulary knowledge tasks and LD measures demonstrate closer relationships along with the increasing IELTS writing scores, especially Lex30 and the PVLT, which increase from weak correlations to moderate and strong. G_Lex shows no significant correlation with LD measures at level 5.5, but shows moderately significant correlations at levels 6.0 and 6.5. The PVLT task scores indicate the most sensitive changes across the three IELTS writing scores. Meanwhile, the results in Table 5.7 show significant

correlations between the three productive vocabulary knowledge tasks and LD measures for all 98 participants. The PVLT shows the strongest positive relationships with lexical diversity measures, followed by Lex30 and G_Lex. The regression analysis results in Table 5.10 show that the PVLT can explain the largest proportion of the variance in LD scores, followed by Lex30 and then G_Lex. The regression analysis results in Table 5.11 show that, along with the improvement of writing proficiency, vocabulary scores can account for a greater proportion of variance in lexical diversity scores.

The second question asked whether participants' vocabulary knowledge increased along with their writing scores. Table 5.12 shows the mean scores for the three productive vocabulary knowledge tasks across the three IELTS writing scores. The mean vocabulary scores in Table 5.12 show that participants with a higher IELTS writing score also have higher scores on their productive vocabulary tasks. This finding shows that participants with higher IELTS writing scores have acquired more productive vocabulary knowledge than participants with lower IELTS writing scores. Thus, increasing one's productive vocabulary knowledge might be an effective way to get a higher IELTS writing score.

Table 5.12

Descriptive Statistics of Vocabulary Tasks of Different IELTS Writing Scores

Vocab tasks	5.5 (n=28)		6.0 (n=49)		6.5 (n=21)	
	Mean	SD	Mean	SD	Mean	SD
Lex30%	20.48	6.00	21.80	8.6	28.41	10.80
G_Lex%	10.33	4.26	10.68	4.78	14.13	8.25
PVLT%	17.02	7.01	26.24	18.35	43.23	22.18

5.4.1 Limitations

A potential limitation of the experiment in the current study concerns the number of participants with higher IELTS writing scores. The participants' IELTS writing scores for the current study range from 5.5 to 6.5. The number of participants whose IELTS writing scores are equal to and over 6.5 is 21, compared with those whose IELTS writing scores fall at 6.0 (n=49) or 5.5 (n=28). Including more participants with higher IELTS writing levels might help to maintain a better balance between the number of participants in each group.

5.4.2 Conclusion

The current chapter has explored whether productive vocabulary knowledge tasks can differentiate between IELTS scores and whether participants' vocabulary knowledge increases along with their IELTS writing scores. I asked all 98 participants to complete three productive vocabulary knowledge tasks (Lex30, G_Lex, and the PVLT) and two IELTS writing topics. The results in the current chapter show that all three productive vocabulary knowledge tasks can differentiate between IELTS writing scores. Participants with higher vocabulary knowledge will achieve a higher IELTS writing score.

All three experimental chapters so far have explored the relationships between vocabulary knowledge and IELTS writing levels/scores from a cross-sectional perspective. As stated at the start of this chapter, increasing vocabulary knowledge leads to more links/networks between words. Participants who acquire more vocabulary knowledge tend to build more links between words than participants who have acquired less vocabulary knowledge. Examining the extent to which vocabulary knowledge changes over time in productive vocabulary knowledge may also be reflected in vocabulary use in writing. Therefore, the next experimental chapter investigates vocabulary development by conducting a longitudinal study and employing the same participants twice to complete the three productive vocabulary tasks and two IELTS writing topics.

Chapter 6: To What Extent Do Productive Vocabulary Knowledge Task Scores and Lexical Diversity Measure Scores Relate Over a Short Study Period?

6.1 Introduction

The previous experimental chapters (chapters 3, 4, and 5) have examined potential relationships between vocabulary knowledge tasks and writing proficiency. To summarise, chapter 3 compared relations between four vocabulary knowledge tasks (the VLT, Lex30, G_Lex, and the PVLT) and written production for L1 Chinese participants (n=29). According to the findings in chapter 3, there were no correlations between the vocabulary knowledge tasks and writing at this participant level. A potential reason for this lack of correlations in chapter 3 may have been the low-level participants' (CEFR=A2) limited language ability in applying their vocabulary knowledge to their written production. This finding motivated the studies reported in chapters 4 and 5. The findings in those following experimental chapters (chapters 4 and 5) showed that three productive vocabulary knowledge tasks (Lex30, G_Lex, and the PVLT) could, to some extent, predict the writing proficiency levels with different degrees of strength. Chapter 4 showed significant correlations between the productive vocabulary knowledge tasks and lexical diversity measures for 91 L1 Japanese participants, whose proficiency ranged from CEFR B1 to C1. Chapter 5 continued this theme by investigating relationships between vocabulary tasks and lexical diversity measures to determine whether such tasks could differentiate between IELTS writing proficiency levels. Chapter 5 investigated 98 (63 L1 Japanese and 35 L1 French) L2 English learners with three productive vocabulary tasks (Lex30, G_Lex, and the PVLT) and two IELTS writing topics. Trained IELTS raters graded the writing samples from band 0 to band 9 based on the IELTS band descriptors (see Appendix C, presenting them in full). The findings in chapter 5 demonstrated that lexical diversity measures can differentiate between IELTS proficiency levels for participants whose writing scores were at IELTS 5.5, 6.0, 6.5 or above. These

findings suggest that, for participants with higher IELTS writing scores, there are closer relationships between productive vocabulary knowledge scores and lexical diversity scores than are found with their lower-scoring participant counterparts (with lower IELTS writing scores). The significant correlations ranged from weak to moderate or strong correlations in chapter 5 across the three productive vocabulary knowledge tasks (Lex30, G_Lex, and the PVLT, respectively), and the lexical diversity measure scores suggested that participants with larger vocabulary knowledge attain higher scores in their IELTS writing.

The three experimental chapters summarised above have investigated potential relationships between vocabulary knowledge tasks and writing proficiency by adopting a cross-sectional, single capture perspective. Because language learners are determined to improve their language knowledge, and because we can see an improvement in the tasks employed in this dissertation, I am keen to explore whether a developmental study might reveal such gains or indications of improvement. There is a shortage of vocabulary knowledge development research in the vocabulary research community. Kremmel and Pellicer-Sánchez (2020) showed that many vocabulary measures could measure vocabulary knowledge development but '…will not easily show gains after short learning or intervention periods' (p. 217). They also call for vocabulary researchers to utilize vocabulary measures in best practice because of a lack of validation evidence. Similarly, Pellicer-Sánchez (2019) mentioned that despite the clear benefits of conducting longitudinal studies of vocabulary knowledge growth, an insufficient number of studies have sought to examine this topic.

Several studies have investigated vocabulary knowledge development and reported a non-linear growth of vocabulary knowledge (Cobb & Horst, 2001; Daller et al., 2013; Fitzpatrick, 2012). Daller et al. (2013) investigated the learning curve and vocabulary development model in a longitudinal study with large empirical data. They used human raters and lexical diversity measures to judge general writing proficiency. Their study investigated

42 participants studying for a bilingual English and Arabic degree. The participants in their study were in a two-year-long English language program before they entered their subject area in university. Their research collected one essay writing sample from each participant every ten teaching weeks, and their final analysis reported 294 writing samples from seven occasions. They used two lexical diversity measures (Guiraud, also known as Root_TTR, and D) and human raters to determine whether vocabulary knowledge develops according to these measures. Their study showed a learning curve with a power function based on human judgement and a structural equation model. Their findings showed participants' writing proficiencies improved in a non-linear learning pattern towards vocabulary acquisition.

Cobb and Horst (2001) used a receptive vocabulary knowledge measure, the Vocabulary Levels Test (VLT; Nation, 1983; Schmitt et al., 2001), in both a pre- and post-test to investigate vocabulary knowledge development of reading ability using online vocabulary learning tools. Their study investigated 33 participants, and the intervention period was 13 weeks in line with the course teaching. The post-test results in their study showed that participants' mean VLT scores improved over the experimental course. Their paper suggested that participants' receptive vocabulary knowledge could improve in a short time, and an online-based vocabulary learning tool was an effective way of examining language learners' vocabulary knowledge.

Fitzpatrick (2012) was the first study to investigate productive vocabulary knowledge growth through a lexical knowledge elicitation task, Lex30. She collected the data longitudinally and investigated whether her participant's vocabulary knowledge developed in both receptive and productive vocabulary knowledge: form and meaning (receptive), written form (productive), word parts (productive), associations (productive), and collocation (productive). One participant, a native speaker of Chinese studying at a university in the UK, joined her study. She used a repeated version of the Lex30 task on six different occasions and

tested the participant over a six-to-eight-week interval in an academic year. Her paper showed that Lex30 effectively elicited the participant's vocabulary knowledge from the perspective of vocabulary knowledge development. The findings in her study showed that the participant's vocabulary knowledge relating to associations, collocations, and derived affixes developed linearly; however, the participant's knowledge of learning individual words developed in a non-linear way.

These earlier studies investigated vocabulary knowledge development based on a single vocabulary test (VLT; or Lex30) (Cobb and Horst, 2001; Fitzpatrick, 2012). Daller et al. (2013) used lexical diversity measures and human ratings to investigate writing proficiency growth. No single study has examined both vocabulary knowledge measures and lexical diversity measures to investigate potential vocabulary knowledge or writing proficiency changes. The current chapter will therefore use the same three vocabulary knowledge measures used in the first three chapters to chart potential vocabulary knowledge changes with multiple lexical diversity measures to track writing proficiency growth.

The current chapter investigates whether participants' vocabulary knowledge and lexical diversity scores can improve over a 12-week pre- and post-test design. I chose 12 weeks as the experiment duration based on Elgort's (2018) paper. Her paper reviewed vocabulary knowledge development studies from 2010 to 2017. She reported that, for these studies, 'treatment and study durations ranged from one-off experimental or class sessions to weeks- and months-long studies' and '50 studies used a pre- and post-test design' (p. 9). Ortega and Iberri-Shea (2005) found that 'little is known about the optimal length of observation for the longitudinal study' and 'decisions about how long is long enough for the longitudinal study of L2 development are implicitly made in SLA research by recourse to biological and institutional time scales' (p. 37). They also pointed out that 'eight weeks seems

to be a favored choice' (p. 32) for the longitudinal investigation of L2 instructional effectiveness.

As the results presented in previous experimental chapters (chapter 3 to chapter 5) show, vocabulary knowledge tasks have significant correlations with lexical diversity scores and can predict and differentiate between IELTS writing proficiencies. Nevertheless, the three experimental chapters did not investigate if the same participants would have an increase in their vocabulary knowledge with a longer duration.

Considering the above gaps in the three cross-sectional experimental chapters of the dissertation, I aim to conduct a longitudinal experiment to track vocabulary knowledge growth for the same participants by focusing on both productive vocabulary measures and lexical diversity measures. The current chapter will investigate whether productive vocabulary knowledge tasks can track potential changes in writing scores by concentrating on improving vocabulary knowledge. Therefore, the research questions for the current experimental chapter are:

RQ1: To what extent do productive vocabulary knowledge task scores and lexical diversity measure scores relate to changes over a short study period?

Predictions: I hypothesize that participants, to some extent, will acquire vocabulary knowledge over the short study period. Participants will use the acquired words in their writing, and the increased vocabulary knowledge should be reflected in their lexical diversity scores. Based on the findings from my previous experimental chapters (chapter 4 and chapter 5), there are positive relationships between vocabulary task scores and lexical diversity scores. Participants who obtain higher vocabulary knowledge scores will also achieve higher IELTS writing scores. The results of productive vocabulary task scores and lexical diversity scores in written production should increase. However, the predictions may depend on participants' proficiency levels and motivation during the vocabulary study.

Considering my findings in chapter 4 and chapter 5, G_Lex shows more stable predictions with lexical diversity scores and IELTS writing scores. I expect G_Lex would be more sensitive in tracking vocabulary knowledge changes than Lex30 and the PVLT. The findings from my previous experimental chapters (the cross-sectional studies) also show that the traditional lexical diversity measure (basic measure) scores have more predictive power for vocabulary knowledge task scores and IELTS writing scores than the more recently devised lexical diversity measures. While the current experimental chapter is a longitudinal study, I expect that the more recently devised lexical diversity measures will show more power in tracking vocabulary knowledge development.

RQ2: To what extent do productive vocabulary knowledge task scores and lexical diversity measure scores correlate over a short study period?

Predictions: The findings in my experimental chapters (3, 4, and 5) show that vocabulary knowledge tasks can predict lexical diversity scores and IELTS writing scores. Considering the participants' combined proficiency levels in the current experimental chapter, they all obtain a single proficiency level (CEFR B1), making it difficult to distinguish their IELTS writing proficiency levels using human raters. However, the findings in my experimental chapters show significant correlations between productive vocabulary knowledge tasks and lexical diversity measures for participants ranging from B1 to C1 levels. G_Lex shows a more stable relationship with lexical diversity measures among these correlation results. Productive vocabulary knowledge tasks, especially G_Lex, will show significant correlations with lexical diversity measures for the current experimental chapter.

However, some factors can also influence the results of the present experiment. First, there is no doubt that the learners themselves (as well as their 'motivation, personality, aptitude, and preferred learning style') have a decisive impact on the effectiveness of various vocabulary learning strategies, as pointed out by Gu (2003). In the current experimental

chapter, participants complete most of their vocabulary learning outside of class time under the encouragement of their language instructors, and the present experiment does not track participants' actual study time. It is unknown whether or not they acquired the assigned words. Second, 'flash card learning typically focuses on initial form-meaning mapping and may not facilitate the learning of other aspects such as collocations, associations, or constraints on use' (Nation, 2013, as cited in Nakata, 2020, p. 314). Conversely, the writing task requires participants to use various vocabulary knowledge beyond the 'form-meaning' scope. Therefore, there might be no correlations between their vocabulary knowledge task scores and lexical diversity measure scores after the participants have finished the vocabulary learning process.

6.2 Study

The current experiment explores vocabulary knowledge and IELTS writing level development for the study's participants over a short study span. The study investigates how vocabulary knowledge development can be tracked through a short-term intervention (approximately 12 weeks). I used the same three productive vocabulary knowledge measures (Lex30, G_Lex, and the PVLT) and the same spectrum of major lexical diversity measures in both the pre-test and the post-test times. Participants from a single language background (L1 Japanese), with similar proficiency levels, took part in the current study. The experiment reported in the current chapter uses the same word unit, the flemma, as described in chapter 4 and chapter 5, for productive vocabulary tasks and lexical diversity measures.

6.2.1 Measures

The current chapter uses two different versions of the three productive vocabulary knowledge tasks (Lex30, G_Lex, and the PVLT), as outlined in section 3.2.1.1 of chapter 3.

The chapter uses multiple lexical diversity measures, Types, D (vocd), HD-D, TTR, Log_TTR, Root_TTR, MSTTR, MAAS, MATTR, MTLD, and MTLD_W, as described in section 3.3.1.2 (an introduction of these lexical diversity measures).

6.2.2 Participants

The experiment conducted in the current chapter reports on a longitudinal study for 51 L1 Japanese participants with the three productive vocabulary tasks (using two different versions) and two IELTS writing samples (with different question prompts at each test time) at the beginning and the end of the study period. I gave all participants the same vocabulary lists to learn for the study period (https://quizlet.com/jp/546414568/ngsl-20-engjap-1001-1100-flash-cards/). The participants were 51 L1 Japanese undergraduates from a university in Japan. They came from different subject majors and were in their first year of university studies. The participants were aged between 18 to 19 years old. I required participants to complete the three vocabulary tasks (Lex30, G_Lex, and the PVLT) and two different IELTS writing topics, one week before the intervention began and immediately after the intervention. Their CEFR levels belonged to B1, as judged by their English language instructor. Their instructor told them not to refer to any materials while responding to the vocabulary tasks or undertaking the writing process. For the pre-test, I asked the participants to finish the three vocabulary tasks (Lex30, G_Lex, and the PVLT) in the first week and the two IELTS writings the following week. As with the post-test, I asked the participants to complete the three vocabulary tasks (the vocabulary tasks presented were different versions from the pre-test versions) and two IELTS writing topics (the writing topics also differed from the pre-test) within two weeks after the vocabulary intervention. The participants in the current chapter gave their consent to join the study, and the process followed the ethical procedures. All participants took the experiment voluntarily and reserved the right to

withdraw. The Research Ethics Committee of the Graduate School of Humanities and Social Sciences, Hiroshima University, has approved this research (approval number: HR-HUM-000804).

To justify the minimum sample size requirement for the current chapter, I conducted a priori power analyses through G*Power (Faul et al., 2007) for a paired samples T-test and a Wilcoxon signed-ranks test. I first conducted the power analysis for the paired samples T-test. To reach a medium effect size (Cohen's d=0.5) and to achieve 80% power (80% chance to detect a significance) for my hypothesis, with the significance value of 0.05 (α err prob), the results showed that a minimum sample size would be 34 for the selected statistical test in G*Power (Means: Difference between two dependent means of matched pairs). I also ran a power analysis for a Wilcoxon signed-ranks test and chose the min ARE (asymptotic relative efficiency) for the parent distribution. To reach a medium effect size (Cohen's d=0.5) and to achieve 80% power (80% chance to detect a significance) for my hypothesis, with the significance value of 0.05 (α err prob), the results showed that a minimum sample size would be 39 for the selected statistical test (Means: Wilcoxon signed-rank test for matched pairs). Therefore, with the participants numbering 51 for the pre-test and post-test, I viewed the sample size as adequate to meet the required sample size by the power analysis.

6.2.3 Methodology

Participants had to complete the three productive vocabulary knowledge tasks (Lex30, G_Lex, and the PVLT) and the two IELTS writing topics, both with different versions, at two different test times (see Appendix F). I told participants the time limitations for each task and required them to submit each task at the allotted time. For Lex30 and G_Lex, I gave the participants 15 minutes for each task, following Fitzpatrick and Clenton (2017). For the PVLT, I required the participants to complete it within 25 minutes, following Edmonds et al.

(2022). Regarding IELTS writing, I gave them 40 minutes for each topic, which is standard practice for IELTS written examinations. One of the writing topics was:

Some people believe that teaching children at home is best for a child's development, while others think that it is important for children to go to school. Discuss the advantages of both methods and give your own opinion. Give reasons for your answer and include any relevant examples from your own knowledge or experience. To track potential vocabulary knowledge growth, the current study concentrated on improving participants' vocabulary knowledge. The primary purpose of the current chapter is to determine which vocabulary knowledge tasks were more sensitive in tracking both vocabulary growth and potential increases of vocabulary knowledge in use across the two test times. Data for this study were collected at two time points: the first in October 2020, and the second in January 2021.

I selected the middle 2K words from the New General Service List (NGSL; Browne, 2014) for participants to learn. First, the NGSL is selected based on extensive quantitative data, a 273 million-word subcorpus from the Cambridge English Corpus (CEC) comprising 2 billion words. Second, the complete NGSL word list (around 2800 words) offers 95.2% coverage of the English examinations in Japan. The whole word list covers 95.2% of the National Centre Test (the national university entrance examination in Japan), and its first 1K words cover 98.1% of the High School Entrance Exam (Browne, 2021). All the participants taking part in the current experiment are in their first-year of university study, and they acquired the first 1K NGSL words during their high school education in Japan. Choosing the NGSL words beyond 1K for the participants as an effective word list is relatively reasonable. Third, the whole experiment design must correspond to the course instruction period length. Selecting all words beyond 1K from the NGSL cannot meet the requirement of the teaching sessions, and I choose the mid-2K NGSL words.

Word cards are one of the traditional ways to learn vocabulary. Nation (2022, p. 402) mentioned that word card learning covers form, meaning, and use aspects of vocabulary knowledge. He also indicated that 'learning from word cards is a way of quickly increasing vocabulary size' (Nation, 2022, p. 407). However, 'vocabulary can be presented with the help of technology' (Mahdi, 2018). Using flashcards through web-based pages, personal laptops, or mobile phones is a vocabulary-learning strategy developed under modern technology. According to Nakata (2011), flashcard software learning is an effective method towards vocabulary learning and 'computer-based flashcards may allow learners to learn more effectively…' (p. 18). He defined flashcard programs as 'software that encourages learners to study L2 vocabulary in a paired-associate format' (Nakata, 2011, p. 17). In other words, the program combines the target words and language learners' first language (L1) explanation of the words. Nakata (2011) presented a detailed evaluation of current flashcard learning programs based on their design features and contribution to vocabulary learning, and Quizlet was recommended as one of the ideal programs for language instructors. That is one reason why I choose Quizlet as the flashcard learning platform for the current study.

In addition, Nakata (2020) offered a comprehensive review of relevant issues concerning flashcard learning. He emphasised that compared to traditional word cards (paper-based), flashcards (computer-based) have superior aspects. First, flashcards platforms, usually based on mathematical algorithms principles, can enable more effective learning, utilising retrieval practice (to remember the previously learnt words) and the spaced review schedule (the time intervals for better memorisation). Second, the flashcards learning platform can deliver various learning forms that traditional word cards cannot; for example, using sound or videos for practising pronunciation, matching the words with their definitions, offering images, and implementing vocabulary quizzes or typing the correct answers more quickly. Third, flashcards platforms can track the learning records of participants'

achievements, and their language instructors can adjust their course design. Fourth, users can create flashcard learning sets based on personal preferences, for instance, inserting preferred images or creating their own vocabulary sets, and share their flashcard sets with other users.

Considering the advantages of modern flashcard learning, I asked the participants to learn around 100 English words weekly using an online flashcard learning platform. I assigned them words through Quizlet (https://quizlet.com/jp/546414568/ngsl-20-engjap-1001-1100-flash-cards/), an online vocabulary learning platform, each week. The participants reviewed the assigned words each week as warm-up activities within class time under the instruction of their language teachers. I gave the participants the NGSL word list to learn, and each week, I assigned the participants 100 words to learn using flashcards with a bilingual version (Japanese translation).

6.2.4 Data Analysis

I used the same data processing procedures for the scores of the three productive vocabulary knowledge tasks reported in chapter 4 and chapter 5 (see section 4.2.4.1 for more information). I also used the same data processing procedures for the IELTS writing samples for the pre-test and post-test in the same way as reported in chapter 4 and chapter 5 (see section 4.2.4.2 for more information). In the current chapter, I chose only the middle 200 English words for the final analysis.

6.3 Results

After checking the normality of all vocabulary tasks scores and lexical diversity scores, the significant scores in a Shapiro-Wilk test for the three lexical diversity measures (t1_Log_TTR, t1_MAAS, and MTLD values for two test times) and one vocabulary task score (the t1_PVLT) are less than 0.05. The current chapter uses the Wilcoxon signed-rank

test for non-parametric variables to compare the significant changes from pre-test to post-test and the paired samples T-test for the parametric variable for the rest of the variables to compare the differences in the mean scores from pre-test to post-test.

The following figures and tables show the results for the three productive vocabulary knowledge tasks' (Lex30, G_Lex, and the PVLT) scores and the lexical diversity scores across the pre-test and post-test (the two test times). For the convenience of data analysis, in the following tables and figures, I mark all pre-test results belonging to test time one as *t1* and all post-test results belonging to test time two as *t2*. As implemented in the previous chapters, I use the percentage scores for the three vocabulary knowledge tasks (see section 3.2.4.1 for the rationale behind using percentage scores).

Table 6.1 shows the descriptive statistics results of the three productive vocabulary knowledge tasks' (Lex30, G_Lex, and the PVLT) scores for the two test times. The results show that Lex30 (mean=29.17, SD=9.5), a vocabulary knowledge task, elicited the largest vocabulary knowledge at test time one, followed by the PVLT (mean=28.85, SD=10.37) and G_Lex (mean=16.29, SD=7.17). As for test time two, the PVLT (mean=26.86, SD=12.33) elicits the largest vocabulary knowledge, followed by Lex30 (mean=24.80, SD=7.37) and G_Lex (mean=23.04, SD=7.16).

Table 6.1

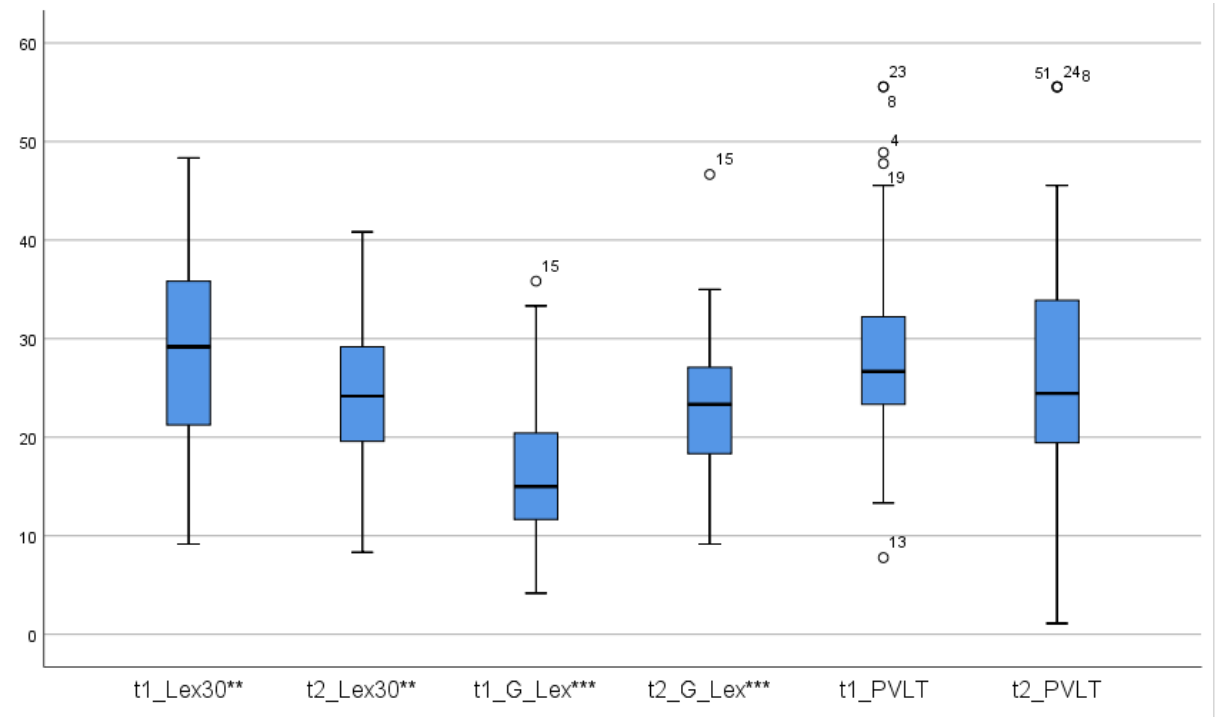
Productive Vocabulary Knowledge (PVK) Tasks Descriptive Statistics

PVK measures (n=51)	t1_Mean	t1_SD	t2_Mean	t2_SD
Lex30%	29.17	9.5	24.80	7.37
G_Lex%	16.29	7.17	23.04	7.16
PVLT%	28.85	10.37	26.86	12.33

Figure 6.1 shows the box plot results for the pre-test and post-tests for the three productive vocabulary tasks (Lex30, G_Lex, and the PVLT). The box plots compare participants' productive vocabulary scores for the two test times. The results in the box plots show that only G_Lex appears to have a developing tendency for participants' vocabulary knowledge by test time two. In contrast, the PVLT and Lex30 show a decreasing tendency for participants' vocabulary knowledge after the intervention. The longer box length (the interquartile range) of Lex30 at test time one and the PVLT at test time two shows that the vocabulary scores are more dispersed than the shorter box length of the other data (t2_Lex30, t1_G_Lex, t2_G_Lex, and t1_PVLT). There is also one outlier for the G_Lex, and several outliers for the PVLT.

Figure 6.1

Pre-test and Post-test Productive Vocabulary Task Scores



Note. **p<0.01; ***p<0.001

Table 6.2 shows the descriptive statistics of the lexical diversity scores in the current pre-test and post-test experiment design. The results in Table 6.2 show that the mean scores of the lexical diversity measures for the post-test are larger than the pre-test lexical diversity scores.

Table 6.2

Descriptive Statistics for Lexical Diversity (LD) Scores

LD measures (n=51)	t1_Mean	t1_SD	t2_Mean	t2_SD
Types	90.29	7.52	94.75	6.45
TTR	0.45	0.04	0.47	0.04
Root_TTR	6.39	0.53	6.70	0.46
Log_TTR	.85	.02	.86	.01
MSTTR	0.70	0.04	0.73	0.03
MAAS	.07	.01	.06	.01
D (*vocd*)	45.84	7.97	51.58	9.48
HD-D	0.75	0.03	0.76	0.02
MTLD	44.24	8.09	51.03	7.32
MTLD-W	43.00	8.16	50.62	6.88
MATTR	0.70	0.04	0.73	0.03

Table 6.3 shows the parametric data's paired-sample t-test and effect size results. The current analysis used a bias-corrected and accelerated (BCa) bootstrapping confidence interval (CI), which is both an accurate bootstrap method for the CI and a means to help

address the outlier concerns, as proposed by Larson-Hall (2015). The bootstrapped BCa CI results in Table 6.3 show that the CIs have not gone through zero, showing statistical differences between the pre-test and post-test. I calculated the Cohen's d (Cohen, 1988) (using means and SDs to calculate) values for the pre-test and post-test comparison of effect size. The means, SDs, and p values (two-tailed) of lexical diversity values in Table 6.3 indicate a significant increase in participant writing levels. Several lexical diversity measures (types, d=0.636; TTR, d=0.573; Root_TTR, d=0.637; MSTTR, d=0.738; D (vocd), d=0.655; and HD-D, d=0.776) indicate medium effect size, and two lexical diversity measures (MTLD_W, d= 0.883; and MATTR, d=0.802) show a large effect size. As with the productive vocabulary knowledge values, the G_Lex scores show a significant increase, $p<.001$ ($p=0.000$), from pre-test (mean=16.291, SD=7.173) to post-test (mean=23.039, SD=7.159), while the Lex30 scores show a significant decrease ($p=0.003$) from pre-test (mean=29.167, SD=9.497) to post-test (mean=24.804, SD=7.367). Lex30 scores (d=0.513) show a medium effect size, and G_Lex scores (d=0.942) show a large effect size.

Table 6.3

Paired-samples T-test and Effect Size Results

N=51	pre-test (t1)		post-test (t2)				BCa 95% Confidence Interval		
	Mean	SD	Mean	SD	t (50)	p	Lower	Upper	Cohen's d
t1_Types - t2_Types	90.294	7.523	94.755	6.451	-4.347	0.000	-6.441	-2.494	0.636
t1_TTR - t2_TTR	0.451	0.038	0.474	0.040	-4.856	0.000	-0.031	-0.014	0.573
t1_Root_TTR - t2_Root_TTR	6.385	0.532	6.701	0.456	-4.355	0.000	-0.458	-0.178	0.637
t1_MSTTR - t2_MSTTR	0.702	0.035	0.725	0.025	-4.246	0.000	-0.033	-0.012	0.738
t1_Dvocd - t2_Dvocd	45.838	7.968	51.575	9.482	-4.485	0.000	-8.197	-3.193	0.655
t1_HD-D - t2_HD-D	0.745	0.027	0.764	0.023	-4.791	0.000	-0.027	-0.012	0.776
t1_MTLD_W - t2_MTLD_W	43.955	8.157	50.617	6.878	-5.108	0.000	-9.131	-4.173	0.883
t1_MATTR - t2_MATTR	0.701	0.035	0.726	0.027	-4.968	0.000	-0.034	-0.016	0.802
t1_Lex30 - t2_Lex30	29.167	9.497	24.804	7.367	3.070	0.003	1.510	7.206	0.513
t1_G_Lex - t2_G_Lex	16.291	7.173	23.039	7.159	-7.311	0.000	-8.660	-4.902	0.942

Note. Bootstrap results are based on 10000 bootstrap sample

Table 6.4 shows the Wilcoxon signed-rank test results for the non-parametric data. The mean scores of lexical diversity measures in the pre-test and post-test show a significant increase in participant writing levels after the intervention (p=0.000) with a medium effect size for Log_TTR scores (d=0.610) and MAAS scores (d=0.601) and large effect size for MTLD scores (d=0.880). The median lexical diversity measure scores increase from the pre-test (Md_{Log_TTR}=0.851, Md_{MAAS}=0.065, Md_{MTLD}=42.508) to the post-test (Md_{Log_TTR}=0.858, Md_{MAAS}=0.06, Md_{MTLD}=50.323). As with the PVLT scores, the pre-test results (mean=28.854, SD=10.366) and post-test results (mean=26.863, SD=12.330), z=-1.668, p>.001 (p=0.095), indicate no significant decrease in participants' PVLT scores, with a small effect size (d=0.174). The median PVLT scores decrease from the pre-test (Md=26.667) to the post-test (Md=24.444).

Table 6.4

Wilcoxon Signed-Rank Test and Effect Size Results

N=51	pre-test (t1)		post-test (t2)		p	z	Md(t1)	Md(t2)	Cohen's d
	Mean	SD	Mean	SD					
t2_Log_TTR - t1_Log_TTR	0.849	0.016	0.858	0.013	0.000	-3.637	0.851	0.858	0.610
t2_MAAS - t1_MAAS	0.066	0.007	0.062	0.006	0.000	-3.637	0.065	0.06	0.601
t2_MTLD - t1_MTLD	44.237	8.092	51.030	7.324	0.000	-4.593	42.508	50.323	0.880
t2_PVLT - t1_PVLT	28.845	10.366	26.863	12.330	0.095	-1.668	26.667	24.444	0.174

Table 6.5 shows the correlations between the three productive vocabulary knowledge task scores and the lexical diversity measure scores. The current analysis uses Pearson's r for parametric data and Spearman's rho for non-parametric data. For the pre-test results, Table 6.5 indicates significant correlations between two productive vocabulary task scores (Lex30 and G_Lex) and lexical diversity scores for the pre-test data. Small but significant correlations ($r=.298^{*}$, $p<0.05$, $n=51$) were found between Lex30 and HD-D scores. Medium significant correlations exist between G_Lex scores and Types scores ($r=.409^{**}$, $p<0.01$, $n=51$), TTR scores ($r=.409^{**}$, $p<0.01$, $n=51$), and Root_TTR scores ($r=.409^{**}$, $p<0.01$, $n=51$). Small but significant correlations exist between G_Lex scores and Log_TTR scores ($rho=.365^{**}$, $p<0.01$, $n=51$), MAAS scores ($rho=-.365^{**}$, $p<0.01$, $n=51$), HD-D scores ($r=.332^{*}$, $p<0.05$, $n=51$), and MATTR scores ($r=.282^{*}$, $p<0.05$, $n=51$). No significant correlations exist between the PVLT scores and lexical diversity scores for the pre-test data. Regarding the post-test results, I have found no significant correlations between the three productive vocabulary knowledge task scores (Lex30, G_Lex, and the PVLT) and the lexical diversity scores.

Table 6.5

Correlations Between Productive Vocabulary Knowledge Scores and Lexical Diversity Scores for Pre-test (t1) and Post-test(t2)

N=51	Types	TTR	Root_TTR	Log_TTR	MSTTR	MAAS	D (*vocd*)	HD-D	MTLD	MTLD-W	MATTR
t1_Lex30	0.269	0.269	0.269	0.232	0.237	-0.232	0.271	.298*	0.195	0.205	0.193
t1_G_lex	.409**	.409**	.409**	.365**	.285*	-.365**	0.267	.332*	0.208	0.204	.282*
t1_PVLT	0.109	0.115	0.108	0.118	0.077	-0.118	0.177	0.171	0.062	0.061	0.046
t2_Lex30	-0.050	0.024	-0.050	-0.063	-0.214	0.063	-0.093	-0.103	-0.143	-0.187	-0.228
t2_G_lex	0.046	0.077	0.046	0.055	-0.044	-0.055	0.068	0.120	-0.081	-0.070	-0.082
t2_PVLT	0.271	0.183	0.271	0.258	0.090	-0.258	0.102	0.183	0.103	0.113	0.092

**.Significant at the 0.01 level (2-tailed). *. Significant at the 0.05 level (2-tailed).

6.4 Discussion

The design of the study reported in the current chapter was to explore whether productive vocabulary knowledge tasks and writing tasks potentially track changes in vocabulary knowledge development over a short-term studying period (12 weeks). Using the same three productive vocabulary knowledge tasks (Lex30, G_Lex, and the PVLT) as those used in chapters 3, 4, and 5, and two IELTS writing tasks, I investigated participants (N=51) at two test times, in a pre-test and post-test design. The introduction to the current chapter highlighted the importance of conducting a longitudinal experiment in vocabulary studies to track potential vocabulary knowledge changes. Adopting a paired samples t-test and a Wilcoxon signed-rank test, the results show that all lexical diversity measures and one productive vocabulary knowledge task (G_Lex) appear to indicate vocabulary knowledge growth for the L1 Japanese participants. The correlations show that productive vocabulary knowledge tasks can, to some extent, predict writing levels only for the pre-test lexical diversity scores (i.e., not for the post-test).

The research questions for the current experimental chapter were set as: RQ1: *To what extent do productive vocabulary knowledge task scores and lexical diversity measure scores relate to changes over a short study period?* and RQ2: *To what extent do productive vocabulary knowledge task scores and lexical diversity measure scores correlate over a short study period?* The following sections discuss each of these research questions in turn.

First, the significant differences between the pre-test and post-test results for the three productive vocabulary knowledge task scores show that G_Lex scores ($mean_{t1\_G\_Lex}$=16.291; $mean_{t2\_G\_Lex}$=23.039; p=0.000) can track the vocabulary growth changes of vocabulary knowledge. In contrast, Lex30 scores indicate significant decreases ($mean_{t1\_Lex30}$=29.167; $mean_{t2\_Lex30}$=24.804; p=0.003) through the intervention, and the PVLT scores did not track any change ($mean_{t1\_PVLT}$=28.845; $mean_{t2\_PVLT}$=26.863; p=0.095). The results in Table 6.3

show significant differences for both G_Lex scores (p=0.000) and Lex30 scores (p=0.003). The statistics of G_Lex scores from the pre-test ($mean_{t1\_G\_Lex}$=16.291, SD=7.173) to the post-test ($mean_{t2\_G\_Lex}$=23.039, SD=7.159) in Table 6.3 show a significant increase in vocabulary knowledge with a large effect size (d=0.942). The statistics of Lex30 scores from the pre-test ($mean_{t1\_Lex30}$=29.167, SD=9.497) to the post-test ($mean_{t2\_Lex30}$=24.804, SD=7.367) show a significant decrease in vocabulary knowledge with medium effect size (d=0.513). The results in Table 6.4 show no significant changes in the PVLT scores (p=0.095) from the pre-test ($mean_{t1\_PVLT}$=28.854, SD=10.366) to post-test ($mean_{t2\_PVLT}$=26.863, SD=12.330) results with a small effect size (d=0.174).

Second, the significant differences between the pre-test and post-test for the lexical diversity measure scores indicate that lexical diversity measures can, to some extent, show increases in vocabulary (assuming greater diversity shows vocabulary usage). Accordingly, the larger lexical diversity scores in the post-test compared with the pre-test results show a growing vocabulary within the IELTS writing samples. The effect sizes for multiple lexical diversity measure scores from pre-test to post-test ranged from middle effect size to large effect size. The large effect size of the lexical diversity measures exists in MTLD_W scores (d=0.883), followed by MTLD scores (d=0.880) and MATTR scores (d=0.802), which suggests that the more recently established lexical diversity measures offer a greater practical application in identifying developmental changes than the traditional lexical diversity measures. I return to this question in my discussion chapter.

Third, the second research question asked to what extent the productive vocabulary knowledge task scores and lexical diversity scores correlate over a short study period. The correlation results in Table 6.5 indicate that productive vocabulary knowledge task scores can only predict vocabulary in use at the pre-test, not the post-test. Regarding the pre-test, Lex30 scores indicate small but significant correlations with HD-D scores ($r=.298^{*}$, $p<0.05$, n=51),

and G_Lex scores indicate small to moderately significant correlations with most lexical diversity measure scores, aside from the three lexical diversity measures of D (*vocd*), MTLD, and MTLD_W. I found no significant correlations between the PVLT scores and LD measure scores for the pre-test.

A potential reason for this may be that embedded vocabulary tests are more effective in predicting reading ability than discrete vocabulary tests (Jeon & Yamashita, 2014). This suggestion that embedded vocabulary tests might predict other skills may also apply to writing. Given the levels of the participants in the current experimental chapter, G_Lex could have a closer relationship with participants' ability to use their vocabulary knowledge compared to Lex30 and the PVLT. No significant correlations exist between the three productive vocabulary knowledge task scores (Lex30, G_Lex, and the PVLT) and the lexical diversity measure scores for the post-test data. However, the results in the previous chapters (chapters 4 and 5) indicate significant correlations between productive vocabulary task scores (Lex30, G_Lex, and the PVLT) and lexical diversity measure scores, suggesting that productive vocabulary knowledge tasks can, to some extent, predict the vocabulary used in the IELTS writing tasks. These significant correlations indicate that participants with higher productive vocabulary knowledge scores also achieved higher lexical diversity scores in their writing. The study reported in this current chapter, however, shows that no significant correlations exist between the three productive vocabulary knowledge task scores and the lexical diversity measure scores in the post-test data. A reason for this lack of correlation may be the length of the intervention and the waning motivation of the participants towards the word-list studies. I return to this question in the following discussion chapter.

6.4.1 Limitations of the Present Study

The study reported in the current chapter is potentially limited by both the single language background (L1 Japanese) of participants and the length of the intervention. First,

the study only includes L1 Japanese participants who are undergraduates at a university in Japan. Their English language proficiency levels are B1. The number of participants is 51. I have not included participants from other language backgrounds with different proficiency levels. The findings might differ if the current chapter included participants from diverse language backgrounds with different language proficiency levels. These factors might be worthy of follow-up studies. Second, the total period of intervention for the current study was 12 weeks, as was the intervention study conducted by Cobb and Horst (2001), but some studies have also examined vocabulary knowledge growth in language learners for a year-long study period (e.g., Daller et al., 2013; Fitzpatrick, 2012). Third, even though I required the participants to learn the words using Quizlet during the out-of-class time, what remains unknown is whether they actually did it and how assiduously.

6.4.2 Conclusion

The study reported in the current chapter has investigated whether productive vocabulary knowledge task scores and lexical diversity measure scores can track vocabulary knowledge growth for 51 L1 Japanese participants. I conducted the study through flashcard learning using an online vocabulary learning platform under the instruction of their language teachers. I required the participants to complete the same three productive vocabulary knowledge tasks (Lex30, G_Lex, and the PVLT) (different versions for the pre-test and the post-test) and two IELTS writing tasks (both on different topics). The t-test results show that participant vocabulary knowledge in writing improved over the short study period. The vocabulary knowledge scores also improved when reported by G_Lex task scores. Correlation results show that the G_Lex task appears to be the most sensitive in tracking vocabulary knowledge improvement, followed by Lex30. In contrast, the PVLT scores did not show changes in the pre-test and post-test results. The findings reported in the current

chapter suggest that intentional vocabulary knowledge learning can, to some extent, improve participants' vocabulary knowledge and their vocabulary used in written production.

Chapter 7: Discussion

7.1 Introduction

I divide the discussion chapter into seven main sections. The first section (section 7.2) summarises the main findings from the four experimental chapters. First, I summarise the primary research questions and restate the hypotheses relating to the experimental chapters. Second, I present a recapitulation of the experiments performed for each chapter, along with the key findings.

The second section (section 7.3) discusses how vocabulary knowledge measures show discrepancies in assessing vocabulary used in written production. First, I emphasise the importance of investigating vocabulary knowledge in use with Bachman and Palmer's (2010) Assessment Use Argument (AUA) theory. Second, I examine different correlations between vocabulary knowledge measures and lexical diversity measures across the four experimental chapters. Third, I assert that vocabulary knowledge measures differ in their embeddedness and the extent to which they can predict writing. Fourth, I review vocabulary knowledge measures that can distinguish different writing levels judged by human raters.

The third section (section 7.4) examines the word count unit selection in the dissertation. First, I deliberate on the significance of selecting appropriate word counting units for my current dissertation. Second, I evaluate the importance of keeping word unit selection consistent in each study for vocabulary knowledge and lexical diversity measures.

The fourth section (section 7.5) discusses how I score vocabulary knowledge measures. First, I scrutinise how to score vocabulary knowledge tasks by comparing them with previous studies. Second, I examine the concurrent validities with previous studies using the same vocabulary tasks.

The fifth section (section 7.6) explores the potential findings relating to lexical diversity measures. First, I examine the strong correlations between lexical diversity

measures. Second, I identify which lexical diversity measures show closer relationships with vocabulary knowledge measures in my cross-sectional chapters. Third, I examine which lexical diversity measures can track vocabulary knowledge changes in IELTS written production over a short study period.

The sixth section (section 7.7) examines acquired vocabulary knowledge from the NGSL lists. First, the section analyses the number of 2K NGSL words acquired by participants through three vocabulary knowledge tasks and writing samples. Second, the section features an individual examination of the acquired words of four randomly selected participants.

The seventh section (section 7.8) discusses the limitations, outlines the implications for future research, and concludes this discussion chapter by recapitulating the main points and highlighting the significance of the findings.

7.2 Experimental Chapter Main Findings

The four experimental chapters from chapter 3 to chapter 6 examined three main questions: (i) the extent to which vocabulary knowledge tasks could predict vocabulary knowledge use in IELTS writing for participants belonging to a single proficiency level (A2) or different CEFR levels (B1, B2, and C1); (ii) how productive vocabulary knowledge scores could distinguish between different IELTS writing levels assessed by human raters; and (iii) the extent to which productive vocabulary task scores and lexical diversity scores could track the changes for participants over a short study period. I hypothesised that vocabulary knowledge was a significant factor in predicting vocabulary in writing (mainly investigated in chapter 3 and chapter 4), and that vocabulary knowledge could distinguish IELTS writing levels (mainly investigated in chapter 5). I employed previously validated vocabulary knowledge tasks to assess participants' vocabulary knowledge and lexical diversity measures

to assess vocabulary in written production. Specifically, I used human raters to judge writing samples based on IELTS writing band descriptors to distinguish participants' IELTS writing levels. Another hypothesis I made was that vocabulary knowledge tasks could track participant vocabulary knowledge growth through a short study period (investigated in chapter 6). I assigned participants the 2K level words from the NGSL word list to learn using Quizlet under the instruction of their English language instructors. With these questions and hypotheses above, I conducted four experiments in my dissertation, and the following paragraphs present a summary of the respective and aggregated main findings.

7.2.1 Main Findings of Experiment 1 in Chapter 3

The first experimental chapter (3) explored the potential relationships between vocabulary measures and L2 written production for participants at the A2 level: a partial replication of Treffers-Daller et al. (2018). Chapter 3 used four vocabulary knowledge measures (Lex30, G_Lex, the PVLT, and the VLT) to assess participants' (29 Chinese undergraduates) vocabulary scores and multiple lexical diversity measures (Types, D (vocd), HD-D, TTR, Log_TTR, Root_TTR, MSTTR, MAAS, MATTR, MTLD, MTLD_W) to evaluate their IELTS writing. Chapter 3 employed lemmas as the word counting unit for vocabulary knowledge and lexical diversity scores to replicate Treffers-Daller et al. (2018) partially.

The findings in chapter 3 showed that there were significant correlations between Lex30 and G_Lex scores, and lexical diversity scores were also positively correlated. However, there were no correlations between vocabulary knowledge task scores and lexical diversity scores. The regression analyses showed that vocabulary knowledge measures could explain a minor proportion of the variance in lexical diversity scores. I concluded that the main reason for the differences in results between my participants and those of Treffers-

Daller et al. was that they differed in their proficiency level in English. The proficiency levels of the participants in chapter 3 correspond to the A2 level, while the proficiency levels of the participants in the Treffers-Daller et al. study ranged from B1 to C2. In addition, the first experiment used four vocabulary tasks and reported the actual vocabulary scores, while the PTE Academic test provided the vocabulary scores for Treffers-Daller et al. study. To remedy this issue of participants' mismatched proficiency levels, I conducted the second experiment (chapter 4).

7.2.2 Main Findings of Experiment 2 in Chapter 4

The second experimental chapter (4) investigated potential relationships between productive vocabulary task (Lex30; G_Lex; and the PVLT) scores and L2 written production for participants at levels B1 to C1 (compared to the A2-level participants in chapter 3). Since Treffers-Daller et al. (2018) study did not employ the flemma as a word counting unit for lexical diversity measures, the second experiment and the subsequent experimental chapters (chapter 5 and chapter 6) utilised the flemma as the word counting unit to address the research gap and accordingly flemmatised the responses in the vocabulary knowledge tasks and writing samples. Considering that a recent published paper (Edmonds et al., 2022) interpreted that the performance of the PVLT task represented receptive vocabulary knowledge, I excluded the VLT task, a typical receptive vocabulary knowledge test, for the second experiment and instead only used three productive vocabulary knowledge tasks. Also, because writing is a productive skill, I wanted to limit the number of vocabulary tasks by concentrating on tasks that elicited productive vocabulary knowledge.

The findings in chapter 4 demonstrated that productive vocabulary task scores could predict lexical diversity scores in written production. The results showed moderately significant correlations between productive vocabulary task scores and lexical diversity

scores. The PVLT scores exhibited closer relationships with lexical diversity scores than either the G_Lex scores or Lex30 scores did. The regression analyses showed that the G_Lex and the PVLT scores could explain more variance in lexical diversity scores than Lex30. Specifically, the G_Lex scores could explain the greatest percentage of the variance in TTR scores (35.7%), and the PVLT scores could explain the second greatest percentage of the variance in TTR scores (33.6%). Only 16% of the variance in Lex30 scores could be accounted for by TTR scores. Considering the findings in the second experimental chapter, I wanted to investigate whether vocabulary knowledge scores could distinguish different IELTS writing scores and track participants' changes in vocabulary and lexical diversity scores over a short study period.

7.2.3 Main Findings of Experiment 3 in Chapter 5

The third experimental chapter (5) examined the extent to which productive vocabulary tasks could differentiate between IELTS writing scores. The third experimental chapter used trained IELTS raters to mark the IELTS writing samples based on the IELTS writing band descriptors. The previous two experiments investigated whether vocabulary knowledge tasks could predict IELTS writing scores for participants of different CEFR levels. As a departure from the first two experiments, the third experiment chapter investigated how vocabulary knowledge task scores could distinguish between different IELTS writing scores. I further divided all participants (n=98) into three groups based on the IELTS raters' judgements. The third experiment chapter used the same vocabulary knowledge tasks and lexical diversity measures as the second. I required the participants in the third experiment to produce two pieces of IELTS writing, and I calculated the mean scores of their writing scores from the human raters' scores and their lexical diversity scores.

The findings in chapter 5 presented correlations and regression analyses between productive vocabulary knowledge tasks and lexical diversity measures. Looking at all participants, the strongest and the most significant correlations and R^2 values existed with the scores between the PVLT and lexical diversity, followed by the scores between Lex30 and lexical diversity, and then the scores between G_Lex and lexical diversity. I divided the participants into three groups. Their vocabulary knowledge scores and lexical diversity scores would increase proportionally along with their writing levels. Participants with higher IELTS writing scores demonstrated closer relationships between vocabulary knowledge task scores and lexical diversity scores, as shown by both correlation and R^2 values. The PVLT task scores showed the strongest significant correlations with MSTTR for participants at the highest proficiency level, and it was also shown that the PVLT scores could explain the largest percentage of variance in MSTTR scores.

7.2.4 Main Findings of Experiment 4 in Chapter 6

The fourth experimental chapter (6) explored the vocabulary knowledge development of the participants (n=51) over a short study time. Because of a shortage of vocabulary knowledge development studies and language learners' determination to improve their language abilities, chapter 6 investigated the possibility of tracking this development using the vocabulary knowledge tasks and lexical diversity measures employed in the current study. I asked the participants to participate in an experiment that included both a pre- and post-test, and I assigned them two versions of three productive vocabulary knowledge measures and two versions of two IELTS writing topics. I asked them weekly to learn the NGSL words from the 2K level. Their proficiency levels corresponded to CEFR B1 learners.

The findings in chapter 6 showed that all lexical diversity measures and one vocabulary knowledge measure (G_Lex) could track the vocabulary development changes

across two testing times. The Lex30 scores showed a significant decrease at testing time 2. However, the PVLT scores could not track significant vocabulary knowledge increase and decrease changes. Lexical diversity scores detected the growing vocabulary knowledge within IELTS writing. The more recently developed three lexical diversity measures (MTLD, MTLD_W, and MATTR) demonstrated more practical applicability for identifying vocabulary knowledge growth changes than the traditional lexical diversity measures.

7.3 Discrepancies in Accessing Vocabulary Knowledge Use in Written Production Demonstrated by Vocabulary Knowledge Measures

7.3.1 The Importance of Investigating Vocabulary Knowledge in Use

This dissertation investigated the role of vocabulary knowledge in evaluating L2 participants' writing. Bachman and Palmer (2010) classified listening, reading, speaking, and writing as 'language use activities' instead of 'language skills' (p. 34). The current discussion section follows Bachman and Palmer's characterisation of writing as a 'language use activity'. Bachman and Palmer (2010) suggested that 'language knowledge can be thought of as a domain of information in memory that is available to the language user for creating and interpreting discourse in language use' (p. 44). In the current dissertation, I treat vocabulary knowledge as a storage of information in language learners' minds that needs to be stimulated in producing their writing. This elicitation should be reflected in vocabulary knowledge tasks and written production responses. Assessing test-takers' vocabulary knowledge can also raise their awareness of their written or spoken production. Test-takers with more vocabulary knowledge also have relatively higher writing or overall language proficiency levels. Bachman and Palmer (2010) stated that 'assessment tasks or texts that we include in the assessment need to be selected with an awareness of what other areas of language knowledge they may evoke' and 'other areas of language knowledge will inevitably be involved in

language assessment performance' (p. 44). Their statements supported the findings in my experimental chapters that vocabulary knowledge tasks which involved different aspects of linguistic knowledge appear to have different correlations and predictive relationships with writing. Based on the findings from my experimental chapters, these different predictive powers and correlations between vocabulary knowledge tasks and writing are also influenced by the proficiency levels of test-takers.

Bachman and Palmer (2010) proposed an Assessment Use Argument (AUA) conceptual framework that provides 'the rationale that we need in order to justify the interpretations and uses we make on the basis of the test takers' performance' (p. 92). They also suggest that when researchers use an AUA to justify the actual consequences, decisions, and interpretations based on the assessment. In the following discussion sections, I make several claims about the inferred main statements made from my experimental chapters based on the data results and elaborate on them:

- Disparities in accessing vocabulary knowledge used in written production as evidenced by vocabulary knowledge measures: vocabulary knowledge measures differ in task features, and the degree of embeddedness.
- Measuring vocabulary knowledge can differentiate levels of proficiency in IELTS writing.
- Selecting appropriate measures of lexical diversity depending on the specific research question or goal, as different measures may have different strengths and limitations. Traditional lexical diversity scores show closer relationships with vocabulary knowledge scores in writing use, while the more recently refined lexical diversity measures show better performance in tracking vocabulary knowledge development in written production.

- G_Lex shows greater power in tracking vocabulary knowledge improvement when compared to PVLT and Lex30 for CEFR B1 participants.
- Using online flashcard learning with 2K NGSL lemma-based word lists is an effective way to improve vocabulary knowledge and vocabulary in writing.

7.3.2 Different Relationships Between Vocabulary Knowledge Measures and Lexical Diversity Measures

As mentioned previously (section 3.3.3, section 4.3, and section 5.3), to examine whether vocabulary knowledge can show some agreement with participants' writing proficiency levels, I run correlation analyses between vocabulary knowledge measures and lexical diversity measures. However, these correlation results were inconsistent across the different chapters, as shown in Table 7.1. In the current section, I attribute the different strengths of correlations between vocabulary knowledge scores and lexical diversity scores to two possible reasons: first, the proficiency levels of participants in my experimental chapters would influence the results of the correlations; second, different vocabulary knowledge measures engage different contexts in assessing vocabulary knowledge.

Specifically, the Lex30 scores and PVLT scores show a closer relationship with the lexical diversity scores for the highest-level participants than the G_Lex scores. However, the G_Lex scores show a closer relationship with most lexical diversity scores for participants at the second highest level. Table 7.1 shows the correlations between these three productive vocabulary knowledge measures and the lexical diversity measures of the first three experimental chapters. The first experimental chapter (3) used a cohort of L1 Chinese participants (n=29) whose CEFR levels were A2. The second experimental chapter (chapter 4) used a group of L1 Japanese participants (n=91) whose CEFR levels range from B1 to C1. The third experimental chapter (5) includes participants (n=98) from two different language

backgrounds (63 L1 Japanese and 35 L1 French) whose IELTS writing levels spanned from 5.5 to ≥6.5. I divided the correlations between productive vocabulary knowledge scores and lexical diversity scores based on the different tasks undertaken in the experimental chapters. The Lex30 scores show strong correlations with participants whose IELTS writing levels ranged from 5.5 to ≥6.5 (chapter 5), followed by the participants whose CEFR levels ranged from B1 to C1 (chapter 4), but then no correlations with the A2 level participants reported in chapter 3. The G_Lex scores show the strongest and the most significant correlations with participants whose CEFR levels ranged from B1 to C1 (chapter 4) except for when using some particular lexical diversity measures (MSTTR, MTLD, MTLD_W, and MATTR). The PVLT scores show the strongest and the most significant correlations with participants whose IELTS writing levels ranged from 5.5 to ≥6.5 (chapter 5) and the second strongest significant correlations for participants in chapter 4, whose CEFR levels ranged from B1 to C1. The Lex30 and PVLT scores show higher agreement with the lexical diversity scores for participants whose writing levels have been judged (IELTS writing levels from band 5.5 to ≥6.5), while the G_Lex scores show a closer relationship with most lexical diversity measures for participants with CEFR levels from B1 to C1. Participants' CEFR levels in chapter 5 (intermediate to high proficiency level participants) are higher than participants' proficiency levels in both chapter 4 (intermediate to advanced level participants) and chapter 3 (pre-intermediate participants). Table 7.1 shows the statistics for these correlations.

Table 7.1

Correlation Results Between Vocabulary Knowledge Scores and Lexical Diversity Scores

		Types	TTR	Root_ TTR	Log_TTR	MSTTR	MAAS	D (*vocd*)	HD-D	MTLD	MTLD_W	MATTR
	chapter3	.03	.07	.059	.093	.066	-.096	.152	.147	.048	.07	.195
Lex30%	chapter4	.345**	.357**	.487**	.356**	.271**	-.362**	.303**	.282**	.313**	.264*	.289**
	chapter5	.603**	.603**	.603**	.602**	.576**	-.602**	.557**	.544**	.596**	.578**	.568**
	chapter3	.086	.082	.076	.063	.116	-.042	.091	.114	.049	.059	.102
G_Lex%	chapter4	.479**	.493**	.487**	.482**	.308**	-.510**	.359**	.356**	.262*	.255*	.278**
	chapter5	.403**	.403**	.404**	.403**	.332**	-.403**	.354**	.353**	.325**	.305**	.300**
	chapter3	-.225	-.221	-.226	-.26	-.305	.218	-.268	-.232	-.377*	-.261	-.298
PVLT%	chapter4	.483**	.503**	.490**	.504**	.355**	-.518**	.367**	.358**	.346**	.347**	.347**
	chapter5	.693**	.693**	.693**	.690**	.636**	-.690**	.661**	.649**	.612**	.633**	.620**

Vocabulary knowledge tasks engage different features, such as embeddedness and contextual knowledge, which would also influence their correlations with lexical diversity measures across different proficiency levels. Vocabulary knowledge tasks with different contextual knowledge would support participants' performance on different tasks. Read (2000) proposed three dimensions of vocabulary assessment: discrete vs embedded, selective vs comprehensive, and context-independent vs context-dependent. Read and Chapelle (2001) summarised the key vocabulary tests based on their features, and I modified their tables by adding the tasks used in my dissertation. Using Read's dimensions, I categorise all vocabulary knowledge measures, including IELTS writing tasks, into the three dimensions, as shown in Table 7.2.

Table 7.2

Design Features of the Five Measures

Measures	Features		
The Vocabulary Levels Test (VLT)	discrete	selective	context-independent
Lex30	embedded	comprehensive	context-dependent
G_Lex	embedded	comprehensive	context-dependent
The Productive Vocabulary Levels Test (PVLT)	embedded	selective	context-dependent
IELTS writing task 2 (academic)	embedded	comprehensive	context-dependent
Lexical diversity measures	embedded	comprehensive	context-dependent

The VLT is 'a good example of discrete test' (Read & Chapelle, 2001, p. 4). The VLT assesses vocabulary knowledge based on selected frequency levels and 'the simple structure of the test items' (p. 5). Because the VLT asks participants to match the vocabulary items with their correct explanation, they do not need access to any context. The VLT is a

conventional way to evaluate participants' vocabulary knowledge at the meaning recognition level.

Lex30 is a task based on word association and has long been used in the research community. It does not test for the specific item, and Lex30 belongs to the comprehensive dimension. Lex30 offers a word context, asking test-takers to write any relevant words when they see the prompt words. Participants have the opportunity to write any words related to the single-word context. In the current dissertation, Lex30 is the task targeting embedded features because participants still need to go through an inference process for the elicited words.

G_Lex and the PVLT offer a sentence context, and they are embedded. G_Lex requires test-takers to write five English words at most to fit each sentence gap, and it tests comprehensive vocabulary knowledge. The PVLT requires test-takers to fill out the pre-determined words by giving the first few letters, and it tests the selected vocabulary knowledge.

The IELTS writing task constitutes part of the IELTS test, and it evaluates test-takers' responses from their written responses. IELTS writing task is highly embedded, requiring test-takers to write the words using a heavy load of linguistic knowledge. IELTS topics offer a topic context to the participants, and the vocabulary knowledge used during the writing process appears to build more vocabulary links/networks.

Lexical diversity measures assess vocabulary knowledge in writing tasks for my current dissertation. The vocabulary features in the writing context can be treated as highly context-dependent, comprehensive, and embedded. Thus, the lexical diversity measures should also encapsulate the features of the writing tasks.

As mentioned, the participants in chapter 5 obtained the highest proficiency levels, the participants in chapter 4 were the second highest, and the participants in chapter 3 were the lowest. As presented in Table 7.1, Lex30 and the PVLT task scores (both embedded and context-dependent) make better predictors of high-level participants than do G_Lex scores. Finding significant correlations between vocabulary knowledge tasks and writing proficiency for A2 level participants is difficult. Both embeddedness (which will be discussed in the following section, 7.3.3) and context engagement are important factors when evaluating vocabulary knowledge in activities related to writing proficiency. In addition, Lex30 cues can activate more events from learners' overall lexical resources and access a bigger capture zone (word knowledge) than G_Lex (see Fitzpatrick & Clenton, 2017, p. 862 for more information about the vocabulary test capture model).

7.3.3 Can Vocabulary Knowledge Measures Differ in Their Embeddedness and the Extent to Which They Can Predict Writing?

Based on the results from my first three experimental chapters, the current discussion compares which vocabulary knowledge tasks make better predictors of writing levels according to their embeddedness features. In this section, I introduce the concept of embeddedness, as presented in Jeon and Yamashita's (2014) paper, to examine the predictive power of vocabulary knowledge task scores on writing scores. Jeon and Yamashita (2014) suggested that embedded vocabulary knowledge tests (in which words appear as part of a reading passage or as part of a sentence) or productive vocabulary tests are better predictors of reading ability than either discrete vocabulary tests (in which test terms are free of context)

or receptive vocabulary tests. Jeon and Yamashita reported that a vocabulary measure is embedded if the required word appears within a sentence or words appear as part of a reading passage. Jeon and Yamashita (2014) suggested, importantly, that embedded productive vocabulary measures show better predictions than receptive or discrete vocabulary measures for reading. Their meta-analysis investigated three vocabulary knowledge characteristics: production vs selection, embedded vs discrete items, and completion vs grammatical judgement tests. Their findings show that productive vocabulary measures show a closer relationship with reading (r=.92) than selection vocabulary measures (r=.74); and that embedded vocabulary tests (r=.92) exhibit greater accuracy than discrete vocabulary tests with reading (r=.71).

To the best of my knowledge, no single study has examined this claim concerning the extent to which embedded vocabulary measures can predict writing. To explore Jeon and Yamashita's implication (that embedded tasks better predict language skills) concerning writing, the vocabulary knowledge measures (Lex30, G_Lex, and the PVLT) used in the current dissertation are both productive and, I suggest, embedded to varying degrees, allowing us to make comparisons.

I consider that all three vocabulary knowledge measures used in my study feature characteristics of embeddedness but varying according to the degree to which they are embedded, as proposed by Jeon and Yamashita (2014), because of the context involved within vocabulary knowledge measures. Lex30, which offers a single-word context (test-takers can write any word they can recall), appears to be the least embedded of the three vocabulary measure tasks. G_Lex (requires test-takers to write five different words for the

sentence context) appears moderately embedded. The PVLT offers the most embedded features because it requires test-takers to complete the gap with one specific word. Besides, I posited that greater lexical knowledge positively correlates with better writing performance. Using vocabulary measures that encompass more writing-related knowledge dimensions can provide a more precise explanation of the L2 variance in writing proficiency.

Chapters 3, 4, and 5 investigated the extent to which vocabulary knowledge measures could predict vocabulary knowledge in use by lexical diversity scores in L2 written production and whether vocabulary knowledge scores could distinguish different writing proficiencies. Raters judged the writing samples to evaluate writing levels for chapter 5. I used three productive vocabulary knowledge tasks (Lex30, G_Lex, and the PVLT) to assess vocabulary knowledge and multiple LD measures to assess writing scores objectively.

One main research question I sought to explore in my current dissertation is whether vocabulary knowledge measures can predict vocabulary knowledge in use by lexical diversity scores in L2 written production. The results showed that all three productive vocabulary knowledge tasks could, to some extent, predict writing scores. The regression analyses results support the use of vocabulary tasks (Lex30, G_Lex, and the PVLT) in predicting vocabulary knowledge in use: For participants in chapter 4 (B1 to C1/intermediate to advanced participants), G_Lex scores can explain the maximum variance in lexical diversity scores (with TTR index) followed by the PVLT and Lex30 task scores. G_Lex scores can explain 35.7% of the variance in TTR scores; the PVLT scores can explain 33.6% of the variance in TTR scores; and Lex30 scores can explain 16% of the variance in TTR scores. As with participants in chapter 5 (IELTS writing scores ranged from 5.5 to higher scores/intermediate

to high proficiency participants), the PVLT scores can explain the maximum variance in lexical diversity scores (with TTR index), followed by Lex30 and G_Lex task scores. The PVLT scores can explain 56.9% of the variance in TTR scores; Lex30 scores can explain 40.9% of the variance in MTLD_W scores; and G_Lex scores can explain equal variance (18.9%) in three lexical diversity scores (Types, TTR, and Root_TTR).

The findings in chapters 4 and 5 show that G_Lex and the PVLT task scores appear to be better predictors of lexical diversity scores than do Lex30 scores. As the G_Lex and the PVLT tests have tasks that are embedded (and I contend Lex30 involves the fewest of these task characteristics), the results appear to support Jeon and Yamashita's (2014) claim (for reading), i.e., that embedded vocabulary tasks make better predictors of writing ability. That G_Lex and the PVLT tasks might both superficially appear to elicit productive vocabulary knowledge needs detailing further. A recent paper by Edmonds et al. (2022) suggests that the PVLT task 'patterns with the measure of receptive vocabulary knowledge' and 'representing receptive vocabulary knowledge' (pp. 8–9). Their study questions whether the PVLT is 'the best choice for concurrent validity studies concerning the assessment of productive vocabulary knowledge' (p. 8). Considering their findings, the current study recommends G_Lex as a better measure of productive vocabulary knowledge for L2 writing.

7.3.4 Can Vocabulary Knowledge Measures Differentiate Levels of Proficiency in IELTS Writing?

One of the primary research questions is to determine to what extent the scores of three productive vocabulary knowledge tasks can distinguish between different writing

proficiency levels. The strengths of correlations and R^2 values between the productive vocabulary task scores and LD scores presented in Table 5.9 and Table 5.11 for participants with different writing scores showed that productive vocabulary knowledge tasks predicted different writing proficiencies. These results show that participants with more vocabulary knowledge acquired higher LD scores in their written production. Because the LD measures can predict writing/language proficiency, our finding of R^2 values suggests that vocabulary knowledge tasks can explain the different percentages of variance in different writing scores. Those participants with higher subjectively rated writing scores achieved higher productive knowledge and LD scores. Table 5.9 shows that the correlations between the three productive vocabulary knowledge task scores and lexical diversity scores ranged from no correlation to moderate and strong correlation in line with writing score increases.

The varying strength of the reported correlations and R^2 values between productive vocabulary knowledge and writing proficiencies might reflect the different degrees of contextual engagement required by the three productive vocabulary knowledge tasks. The PVLT requires the greatest attention to context, followed by G_Lex and the Lex30 task. Edmonds et al. (2022) suggested that these three productive vocabulary knowledge tasks (Lex30, G_Lex, and the PVLT) can be viewed according to Nation's (2013) vocabulary knowledge dimensions to support my case regarding contextual engagement, and, I suggest, also the extent to which the tasks are embedded. I, therefore, added writing (Table 7.3) to highlight the different aspects of knowledge that each task tapped.

Table 7.3

Dimensions of Vocabulary Knowledge with Three Vocabulary Tasks and Writing (a Revised Version)

				Lex30	G_Lex	PVLT	Writing
Form	spoken	R	What does the word sound like?				
		P	How is the word pronounced?				
	written	R	What does the word look like?			✓	
		P	How is the word written and spelled?	✓	✓	✓	✓
	word parts	R	What parts are recognisable in this word?				
		P	What word parts are needed to express the meaning?				✓
Meaning	form and meaning	R	What meaning does this word form signal?				
		P	What word form can be used to express this meaning?	✓	✓	✓	✓
	concept and referents	R	What is included in the concept?				
		P	What items can the concept refer to?				✓
	associations	R	What other words does this make us think of?				
		P	What other words could we use instead of this one?	✓	✓		✓
Use	grammatical functions	R	In what patterns does the word occur?			✓	✓
		P	In what patterns must we use this word?			✓	✓
	collocations	R	What words or types of words occur with this one?		✓	✓	✓
		P	What words or types of words must we use with this word?				✓
	constraints on use	R	Where, when, and how often would we expect to meet this word?				
		P	Where, when, and how often can we use this word?			✓	✓

Note. R = receptive knowledge, P = productive knowledge. Reprinted from "Exploring the construct validity of tests used to assess L2 productive vocabulary knowledge," by Edmonds et al., 2022, *System*, *108*, 102855, p. 4 (https://doi.org/10.1016/j.system.2022.102855).

I suggest the PVLT scores more closely relate to the writing construct for participants scoring at a higher band (as supported by the aspects shown in Table 7.3). Language learners with a higher proficiency level and a larger vocabulary better understand the vocabulary used in context and perform better in writing tasks. Their scores on vocabulary and writing more accurately show their language proficiency. I can interpret the less proficient learners' writing scores via their lower productive vocabulary scores, as well as the relative strength of the correlation between their writing scores and productive vocabulary scores. I propose that this multi-task approach shows emerging networks across my participant population. As Milton (2013) explained:

> Once a meaning is attached to that form and some idea is gained as to how the word can be used, then it develops links with other words and begins to network and it does not matter whether these are grammatical or associational or collocational links. (p. 61)

Higher-level participants have greater vocabulary knowledge than lower-level participants, and they can more easily build links with other words and further put their vocabulary knowledge into writing use. For all participant proficiencies, I can interpret such knowledge as emerging vocabulary knowledge in their productive vocabulary task scores. When taken together, the productive vocabulary knowledge task scores appear to show emerging vocabulary knowledge (concerning the writing scores) across the range of language proficiencies. In terms of the current study data, Table 5.9 shows that for participants at writing level 5.5, no correlation exists between productive vocabulary knowledge task and

LD scores; for participants at level 6.0, I find moderate correlations between vocabulary knowledge task (the PVLT shows a closer relationship with LD than Lex30) and LD scores; and, for participants at level 6.5 or above, there are strong correlations between two productive vocabulary knowledge tasks (the PVLT scores show the strongest correlation with LD, followed by Lex30 and G_Lex) and LD scores. This varying correlation strength, along with R^2 values in Table 5.11, between the three productive vocabulary knowledge tasks and the LD scores, accounts for our participant writing proficiency: higher R^2 values show our participants' writing proficiency better than lower R^2 values. The PVLT has the highest R^2 values in distinguishing between writing proficiency because it involves more vocabulary knowledge relating to writing proficiency compared to Lex30 and G_Lex. However, the PVLT shows receptive vocabulary knowledge features based on Nation's dimensions of vocabulary knowledge (as presented in Table 7.3). Also, the findings show G_Lex demonstrates more stability in distinguishing test-takers at two different writing levels (IELTS writing scores of 6.0 and equal to or over 6.5), but Lex30 has higher R^2 values than G_Lex. I can attribute the cause of this to Lex30's status as both a productive and somewhat embedded vocabulary measure. In addition, as proposed by Fitzpatrick and Clenton (2017), Lex30, using 30 cue words to prompt the responses, can activate a higher proportion of infrequent vocabulary knowledge than can G_Lex. G_Lex uses 24 sentences in context, and test-takers have to write the words semantically and grammatically correctly, which means they cannot activate the responses based on word forms and references or L1 words. Thus, it is plausible that Lex30 has a greater capacity to facilitate greater quality and quantity of vocabulary knowledge (than G_Lex) and is a more precise predictor of writing proficiencies.

Lex30 demonstrates more power in predicting writing levels than G_Lex. My claim that Lex30 is a more useful measure of productive vocabulary knowledge may only apply to participants at these specific levels and L1 backgrounds when participating in the current study. I do not know if this represents or applies to all levels of participants.

7.4 Word Counting Unit Selection

In this section, I discuss how the selection of word counting units matters to the current dissertation for both the responses from the vocabulary knowledge tasks and lexical diversity scores for writing samples. Specifically, I state the importance of selecting an appropriate word counting unit for the current research (morphological standards; and the words used for the calculations); and the criticality of keeping word counting units consistent in one study for both vocabulary knowledge measures and lexical diversity measures.

As indicated by Nagy and Scott (2000), both context and morphology (word parts) are 'the two major sources of information immediately available to a reader who comes across a new word' (p. 275). As discussed in section 7.3.2, context awareness plays a significant role in activating vocabulary knowledge. Read (2000) suggested that identifying the word units is 'an important step in research on vocabulary size' (p. 85).

Studies show that vocabulary knowledge contributes to language proficiency (Milton, 2013; Qian & Lin, 2019). Such studies also show that morphological knowledge (i.e., of inflected, derived, and base forms) is essential to vocabulary knowledge. Once we consider the knowledge of morphology an integral aspect of vocabulary knowledge, we need to acknowledge the *word family* (Bauer & Nation, 1993) as a counting unit with different

morphological levels. Creating the word family levels also implies proficiency-related knowledge (Nation, 2021) if, for instance, low-level learners have not acquired comprehensive word family knowledge because of their language proficiency. A more practical comparison might exist between *lemmas* and word families, with the lemma as a part of the word family levels. Bauer and Nation's second word family level can be taken as the lemma level (a headword and its inflected forms), with a *flemma* comprising a headword and its inflected forms comprising different parts of speech.

There has been support in the recent literature for using the lemma or flemma as a suitable word unit for English language learners (Brown et al., 2022; McLean, 2018; Stoeckel et al., 2020). These studies call attention to the widespread acceptance of word family levels as a word unit in language learning, teaching, and vocabulary tests and suggest that any estimation of word part knowledge depends on how 'words' are counted. McLean's (2018) article questioned the validity of using word families as a word counting unit. His article, reviewed in chapter 2, considered a potential gap in word family-based research and focused on participants' ability to comprehend inflected and derived forms. Investigating a cohort of L1 Japanese learners at different proficiency levels, his findings highlight a lack of knowledge in all participants regarding inflectional and derivational forms. McLean thus concluded that the flemma is a more appropriate word unit than the word family for this specific group of language learners.

We can count words in several ways, which can vary according to the purpose of the count and researcher preference. What follows is a brief introduction to the key terms. Word families comprise seven different levels of affixes based on eight level-ordering criteria and

contain a headword plus its inflected and derived forms (Bauer & Nation, 1993). McLean's article uses 'WF6' to refer to the word family levels to level 6, based on Bauer and Nation's word family criteria. Lemmas comprise the base word and its inflected forms with the same part of speech (POS); these include the plural, third person singular, present tense, past tense, past participle, *-ing*, comparative, superlative, and possessive forms. The difference between lemmas and flemmas is that flemmas treat the words in their inflected forms with different POS as the same word. Table 7.4 below summarises the key terms, taking the word *abstract* as an example.

Table 7.4

Key Terms and Their Explanations

Key Term	Explanation
Word unit	The lexical unit comprising a 'word'. The most common of these include tokens (the running words in a text), word types, word families, lemma, flemma, and multiword units.
Word Type	The occurrence of unique words in a text would be counted as different words.
Word Family	Seven different levels were proposed by Bauer and Nation (1993). Word families consist of a headword with its inflected forms and the most derived forms. If we use word family as a word unit, the inflected forms of abstract (*abstract*, *abstracts*, *abstracting*, *abstracted*) and derived forms of abstract (*abstractedly*, *abstractly*, *abstractness*, *abstraction*, *abstractions*) would all be counted as the same word. Word family count assumes that learners have the knowledge of inflected forms and derived forms of the words.
WF6	The term WF6 is used in McLean's (2018) article, which excludes level 7 of Bauer and Nation (1993).
Lemma	Lemma means a headword with its inflected forms of the same part of speech. If we use lemma as a word unit, the adjective *abstract*, noun *abstract/abstracts*, and verb *abstract/abstracts/abstracted/abstracting* would be counted as three different words. Lemma count assumes that learners have the word knowledge of inflected forms but do not have the part-of-speech knowledge.
Flemma	Flemmas are similar to lemmas, but do not distinguish part-of-speech of words. If we use flemma as a word unit, the word *abstract* the adjective (*abstract*), noun (*abstract/abstracts*), and verb (*abstract/abstracts/abstracted/abstracting*) would all be counted as one word. A flemma count assumes that learners have the word knowledge of inflected forms and can distinguish the part of speech of words.

A number of recent papers (Brown et al., 2020; Kremmel & Schmitt, 2016; McLean, 2018; Stoeckel et al., 2020; Treffers-Daller et al., 2018) highlight how different operationalisations of what constitutes a 'word' impact the measures of learners' vocabulary knowledge. In second language writing, in particular, the vocabulary used in learner texts is essential to evaluating language ability. However, before making any quantitative analyses of this vocabulary, studies must ensure the use of appropriate and consistent word counting units.

One means of examining word counting is through the lens of lexical diversity measures. Lexical diversity (LD) measures the variety of word knowledge exhibited in speaking or writing and is used in many assessment tools to predict learner proficiency levels. In chapter 2, I reviewed a recent study (Jarvis & Hashimoto, 2021) investigating three LD measures (MTLD, MTLD-W, and MATTR) using five different word unit operationalisations. However, the results of Jarvis and Hashimoto's study could not distinguish which word counting unit is more suitable for language learners than the others. Thus, this issue of word counting remains a topic of debate, and further research and validation in this area are necessary.

Considering most participants in the current dissertation are from two different language backgrounds (L1 Japanese and L1 Chinese), I take flemma as a word counting unit in chapters 4, 5 and 6. Conversely, considering the participants' proficiency levels in chapter 3 (pre-intermediate participants/CEFR=A2), I assume the participants in this level lack the ability to distinguish the part of speech and thus use lemma as the word counting unit.

Regarding the determination of the final vocabulary for calculating vocabulary tasks and writing responses, word counting pertains to deciding which words are included in the final calculation for lexical diversity measures, as well as which words should be considered in the final calculation for the three vocabulary knowledge tasks. Lexical diversity computational software treats all words appearing in texts as different types if researchers do not define which words should finally be calculated. However, the responses in vocabulary knowledge tasks, such as Lex30 and G_Lex, may include numerous responses that do not fully reflect participants' actual vocabulary abilities, including proper nouns, names of brands/foods/cities, and abbreviations. Therefore, data-cleaning procedures are necessary to ensure that the calculated vocabulary reflects participants' actual lexical ability. Including words that do not reflect participants' actual vocabulary ability in the text would increase/decrease the lexical diversity scores. This could lead to inaccurate evaluations about a study's vocabulary scores based on the obtained scores. Read (2000) also pointed out that 'apart from the problem of distinguishing base and derived words, researchers have to decide how to deal with homographs, abbreviations, proper nouns, compound words, idioms, and other multiword units' (p. 85). In my current dissertation, I have conducted the data cleaning and computation processes for both the responses to vocabulary knowledge tasks and the lexical diversity measures to ensure accurate calculations and valid results (see section 4.2.4 for a detailed explanation of the data-cleaning process and data analysis procedures).

In addition, maintaining consistent word counting methods is another critical issue for the current study. Nation (2021) noted that participants' ability to recognise words of different morphological forms relates to their general language proficiency levels and lexical

ability. During the calculation process, studies must clearly and accurately state what kind of word counting unit is used in their research and why.

The current research mainly focuses on the role of vocabulary knowledge in predicting vocabulary knowledge in use in writing through lexical diversity scores. Additionally, it aims to investigate whether the scores from measures of vocabulary knowledge can predict writing proficiency levels. Both the vocabulary knowledge and lexical diversity measures assess aspects of vocabulary and enable a comparison of lexical ability among participants. It is essential to maintain consistency in the word units used both in responses from vocabulary knowledge tasks and their writing production. There are four experimental chapters in the current dissertation; in the first experimental chapter (chapter 3), lemma as a word counting unit was applied for both the responses from vocabulary knowledge tasks and lexical diversity measures because the participants in this experiment were CEFR A2 level, and, based on the finding in McLean (2018), participants at this level cannot distinguish between inflected forms of some common words. Thus, it was necessary to consider participants' proficiency levels in chapter 3 in selecting the appropriate word counting unit. However, I use flemma as a word counting unit for the rest of the experimental chapters (chapter 4, chapter 5, and chapter 6) since the levels of the participants in these studies is such that they already can distinguish inflected forms and distinguish words of different part-of-speech.

Moreover, as presented in Table 7.4, if one study uses a higher level of word family for the participants in the research, it assumes that the participants have acquired the necessary language ability and proficiency for that specific level of morphological

knowledge. Higher word family levels could overestimate participants' ability to use their acquired morphological knowledge. Lower word family levels could underestimate participants' ability to produce morphological knowledge. Any biases in estimating participants' abilities could cause miscalculations and misrepresentations of their vocabulary knowledge and lexical diversity scores. If one study uses a higher word family level in calculating lexical diversity scores, it causes an overestimation of their vocabulary knowledge. This would result in lower lexical diversity scores in participants' written production because higher word family levels reduce the number of different types produced. Or, if one study uses a lower word family level to compute lexical diversity scores, it would lead to an underestimation of participants' vocabulary knowledge. This would cause the calculation results to show higher lexical diversity scores because lower word family levels would increase the production of different types of words.

7.5 Scoring the Vocabulary Knowledge Measures

Section 7.4 has discussed how word counting unit selection influences vocabulary knowledge scores. This section discusses how I scored the vocabulary knowledge tasks in my current dissertation and the similarities to and differences from previous studies. In addition, to verify whether the scoring standards show concurrent validities, I compare the correlation results in my experimental chapters with previous studies.

Meara and Fitzpatrick (2000) used the *lemma* standard to score Lex30: 'Each of the responses was lemmatised so that inflectional suffixes (plural forms, past tenses, comparatives, etc.) and frequent regular derivational suffixes (-able, -ly, etc.) were counted as

examples of base-forms of these words' (p. 23). Their lemma standard corresponds to the level 2 and 3 word family levels in Bauer and Nation (1993). The responses beyond these two word family levels are treated as different words. The Lex30 scores are the responses beyond level 0 (high frequency structure words, proper names, and numbers) and level 1 (the 1,000 most frequent content words). Their study awarded one point for all the responses outside level 0 and level 1, and the final Lex30 scores were calculated by adding all points together. In addition, Fitzpatrick and Clenton (2017) used WebVP (https://www.lextutor.ca/) to score the responses into frequency levels for Lex30 and G_Lex: the first thousand words; the second thousand words; the academic word list (AWL); and off-list words. They counted the responses beyond 1K (the first thousand words) and used percentage scores.

Edmonds et al. (2022) investigated the construct validity of productive vocabulary knowledge tasks through four tests: Lex30, G_Lex, the PVLT, and the VLT. Their scoring process also calculated the words beyond 1K and the percentage scores of the four tests. They excluded any words that belonged to a function word, proper nouns, or numbers, and awarded each response one point which met the previous criteria. Their paper also pointed out that even though both the PVLT and the VLT tasks claimed to test words from the 2K word families, there were still five words in the PVLT (90 items in total) and 11 words in the VLT (150 items in total) belonging to the 1K word family level. Their research thus excluded the 1K items in the PVLT and the VLT.

The current dissertation mainly investigated three productive vocabulary tasks (Lex30, G_Lex, and the PVLT) from chapters 4 to 6, three productive vocabulary tasks and one receptive vocabulary task (the VLT) in chapter 3, and IELTS writing tasks for all

chapters. The four vocabulary tests used in the current dissertation have different maximum scores. Lex30 and G_Lex require test-takers to produce 120 items maximum, while the PVLT requires 90 items, and the VLT requires 150 items. To make the scores easy to compare and considering that the first thousand words (the most frequent words) cannot represent participants' vocabulary ability, the current dissertation, following Edmonds et al. (2022), uses percentage scores and excludes the outlying responses that belong to the 1K band in the PVLT and the VLT. I also calculated the vocabulary beyond the 1K level using Nation's base word lists, based on the BNC-COCA corpora. I awarded one point to the responses beyond 1K word families. As a departure from Edmonds et al. (2022), and considering the significance of keeping word counting unit consistent for one study, I conducted the same word unit counting and data cleaning procedure for vocabulary tasks as with IELTS writing samples. I conducted the lemma process for Lex30 and G_Lex in chapter 3 as in Fitzpatrick and Clenton (2017) and the flemma process and data-cleaning procedure for Lex30 and G_Lex tasks in my experimental chapters 4, 5, and 6. Because the PVLT and the VLT tasks were created based on the highest word family levels as classified by Bauer and Nation (1993), no words appear in lemma/flemma features. In addition, their characteristics (the PVLT and the VLT require participants to write/select from pre-determined words) also indicate no repetition in their response, unlike Lex30 and G_Lex. For those latter two, participants can write any words they can think of, and there are no pre-determined words, and the responses in Lex30 and G_Lex thus often show features of repetition (before or after the lemma/flemma process), abbreviations, unknown words, and names of persons/countries/cities. In short, it is a necessary step to conduct the

lemma/flemma process for Lex30 and G_Lex, but there is no need to conduct the same lemma/flemma process for the PVLT and the VLT.

To explore whether the vocabulary knowledge tasks in my experimental chapters show concurrent validity, I summarise and compare the significant correlation results with the main studies that have validated Lex30, G_Lex, and the PVLT. Table 7.5 compares the correlation results from my experimental chapters with those reported in previous studies. The findings in my experimental chapters align with published studies' results, except for the A2 level learners in chapter 3. The Lex30 and G_Lex task scores showed moderate correlations for participants in chapters 3, 4, and 5 of my experiment, similar to the results in Edmonds et al. (2022). Fitzpatrick and Clenton (2017) also found significant correlations between Lex30 and G_Lex task scores at .645. Chapters 4 and 5 showed significant correlations between Lex30 and the PVLT task scores, but chapter 3 did not show any for pre-intermediate participants. The correlations between the three vocabulary knowledge task scores (Lex30, G_Lex, and PVLT) reported in chapters 4 and 5 were similar to those reported in Edmonds et al. (2022). Fitzpatrick (2007) and Fitzpatrick and Meara (2004) found moderate significant correlations between Lex30 and the PVLT task scores at .504, while Walters (2012) reported strong significant correlations at .772. G_Lex and the PVLT task scores showed a lack of correlations in chapter 3, moderately significant correlations in chapter 5 at .476, and strongly significant correlations in chapter 4 at .671. Edmonds et al. (2022) reported significant correlations between G_Lex and the PVLT task scores at .527. In short, there were no correlations between the PVLT and Lex30 task scores or the PVLT and G_Lex task scores for pre-intermediate level participants (CEFR=A2). In summary, the

correlation results in my experimental chapters (except for the pre-intermediate level participants) support previous studies using correlation analysis to demonstrate concurrent validity for vocabulary knowledge measures.

Table 7.5

Comparing Correlation Results Between Vocabulary Tasks with the Previous Studies

Studies	Background	Levels	N	Correlation results		
				Lex30&G_Lex	Lex30&PVLT	G_Lex&PVLT
Chapter 3 (Table 3.4)	L1 Chinese	A2 (pre-intermediate)	29	.528**	-.221	.074
Chapter 4 (Table 4.2)	L1 Japanese	B1 to C1 (intermediate to advanced)	91	.590**	.592**	.671**
Chapter 5 (Table 5.4)	L1 Japanese & L1 French	IELTS writing levels from 5.5 to 6.5 or higher (intermediate to high proficiency)	98	.581**	.689**	.476**
Edmonds et al. (2022)	L1 French	highly proficient learners	100	.569	.616	.527
Fitzpatrick and Clenton (2017)	L1 Japanese	pre-intermediate to intermediate	100	.645		
Walters (2012)	L1 Turkish	high-beginning, intermediate, and advanced	87		.772	
Fitzpatrick (2007)	L1 Chinese	intermediate level to advanced	55		.504	
Fitzpatrick and Meara (2004)	L1 Chinese	intermediate level to advanced	55		.504	

7.6 How Lexical Diversity Measures Correlate with Vocabulary Measures and Their Capacity to Track Vocabulary Knowledge Changes

This section discusses which lexical diversity measures show a relatively higher agreement or predictivity with vocabulary knowledge measures across participants of different proficiency levels and which lexical diversity measures show greater ability in tracking vocabulary knowledge changes in writing for participants over a short study period. Before discussing the two questions, I present the rationale behind the strongly or perfectly significant correlations between lexical diversity measures and discuss the current measures measuring part of its construct, as indicated by Jarvis (2013a).

The findings in my experimental chapters show strong and significant correlations among lexical diversity measures (see Table 3.7, Table 4.4, and Table 5.6 for the results of the correlations among lexical diversity measures). One reason for the strong correlation between lexical diversity measures is that the currently developed indices cannot capture the whole construct of lexical diversity. Jarvis (2013a, 2017) explained that lexical diversity measures included seven internal dimensions: variability, volume, abundance, evenness, rarity, dispersion, and disparity. The lexical diversity measures developed so far can only assess the first three dimensions of the lexical diversity construct: variability, volume, and abundance (Jarvis, 2017; Kyle et al., 2021). In experimental chapter 5, lexical diversity scores appear to be perfect correlations (r=1.000**) among five indices: Types, TTR, Root_TTR, Log_TTR, and MAAS. The current dissertation includes all lexical diversity measures developed so far (11 indices). However, the strong correlations reported in my experimental chapters indicate that the lexical diversity indices are measuring the same phenomena, which

implies the construct problems with lexical diversity measures that are still under development (Jarvis, 2017). One solution for future studies is to select several lexical diversity measures, including simple and sophisticated ones, to reduce the computation load without including all lexical diversity indices.

Meanwhile, Jarvis and Hashimoto (2021) emphasised several factors that could influence the results of lexical diversity scores, such as text length, a different window size selection, the accuracy of tagging tools, and the choice of word counting unit. In a recent study, Treffers-Daller et al. (2022) investigated the oral vocabulary ability of Indian primary school children to estimate participants' reading ability to receive English medium instruction (EMI). Their study used two lexical diversity measures: MATTR and the Guiraud index (also called Root_TTR; types/square root of tokens) for the lemmatised texts. Their study set the window size of MATTR to 16 words because two participants in their study produced fewer than 16 words in a story-retelling task. In future studies, researchers can compare lexical diversity scores using window sizes that are similar to or different from those used by Treffers-Daller et al. (2022).

The findings in the current dissertation indicate that lexical diversity scores have varying correlations with vocabulary knowledge measures. Traditional lexical diversity scores (Types, TTR, Root_TTR, Log_TTR, MAAS, and MSTTR) exhibit closer relationships with vocabulary knowledge scores. Although more recently devised lexical diversity measures (MTLD_W, MTLD, and MATTR) have moderate correlations with vocabulary knowledge scores, they demonstrate a large effect size for tracking vocabulary knowledge growth compared to traditional lexical diversity measures.

For the pre-intermediate participants in chapter 3, no significant correlations have been found with the lexical diversity measures. Vocabulary knowledge scores could explain minor, negligible variance in lexical diversity scores. Regarding the intermediate to advanced participants in chapter 4, Root_TTR shows closer relationships among 11 lexical diversity indices with Lex30 scores (r=.487); whereas MAAS index shows a closer relationship with G_Lex scores (r=-.510) and the PVLT scores (r=-.518) than do other lexical diversity measures. As with the intermediate to high proficiency participants in chapter 5, Types, TTR, and Root_TTR show equally closer relationships with Lex30 scores (r=.603); Root_TTR shows a closer relationship with G_Lex scores (r=.404); and three simple lexical diversity measures (Types, TTR, and Root_TTR) show equally closer relationships with the PVLT scores (r=.693). I further investigated, using vocabulary knowledge task scores to differentiate between IELTS writing levels. The findings in Table 5.9 show that for participants whose IELTS writing scores were at 6.0, Root_TTR shows the strongest and most significant correlations with Lex30 scores (r=.584); three lexical diversity measures (Types, TTR, and Root_TTR) show equally the strongest and most significant correlations with G_Lex scores (r=.527); and TTR shows the strongest and most significant correlations with the PVLT scores (r=.655). As with participants whose IELTS writing scores were at ≥6.5, Log_TTR and MAAS scores show equally the strongest and the most significant correlations with Lex30 scores (r=.748); MSTTR shows the strongest and the most significant correlations with G_Lex scores (r=.540), and MSTTR shows the strongest and the most significant correlations with the PVLT scores (r=.798).

Further, chapter 6 investigated whether lexical diversity measures and vocabulary knowledge measures can track the changes for participants over a short study period. The findings in chapter 6 show that all 11 lexical diversity indices can track vocabulary knowledge growth and show medium to large effect size according to Cohen's d effect size (medium=0.5; large=0.8) (see Table 6.3 and Table 6.4 for more information). MTLD_W shows the largest effect size among all lexical diversity measures (d=0.883), followed by MTLD (d=0.880), and then MATTR (d=0.802).

In summary, the traditional lexical diversity measures exhibit closer relationships with vocabulary knowledge measures, while the more recently devised lexical diversity measures show a larger effect size in detecting vocabulary knowledge development. First, the findings emphasise the significance of selecting appropriate lexical diversity measures based on the research aims. The traditional lexical diversity measures accurately predict participants' overall vocabulary knowledge. We can use traditional lexical diversity measures to assess participants' vocabulary knowledge at a specific level. The more recently established lexical diversity measures are more sensitive to tracking vocabulary knowledge growth. We can use more recently established lexical diversity measures to evaluate intervention studies' effectiveness and detect language learning development over time. Second, the findings also inspire further exploration of lexical diversity measures since the more recently established ones offer unique sensitivity in my interventional study.

7.7 Examining Vocabulary Knowledge Acquired From the NGSL Vocabulary Lists

Chapter 6 examined the vocabulary knowledge growth of participants over a short study time (12 weeks). The participants were assigned the second thousand NGSL, a bilingual version (English word lists with Japanese translation), to learn through the online platform Quizlet each week. The participants were required to finish three vocabulary knowledge tasks and two writing topics before the intervention study began and after the intervention ended (different versions of both vocabulary knowledge tasks and writing topics were used for the pre-test and post-test). The findings in chapter 6 indicated that vocabulary knowledge tasks and lexical diversity measures could track vocabulary knowledge improvement. Meanwhile, I also ran correlation analyses for two testing times between vocabulary knowledge tasks and lexical diversity measures. The findings showed that significant correlations only existed in the pre-test, with no significant correlations existing in the post-test. Possible reasons for the lack of correlation in the post-test may be the length of intervention and participants' waning motivation to use Quizlet to acquire words.

Despite the lack of correlation between vocabulary knowledge tasks and lexical diversity measures for the post-test data, participants who engaged in the entire learning process still have opportunities to acquire vocabulary knowledge through the intervention word lists. I have examined the extent to which the participants acquired the NGSL words in the 2K level.

To examine how many 2K level words in the NGSL were acquired by the participants, I matched the 2K NGSL word lists with vocabulary knowledge tasks and writing samples separately at two different testing times. To ensure the matched words demonstrate participants' vocabulary knowledge at testing time 2, I excluded the same 2K NGSL words

that appeared at both testing time 1 (before the intervention study) and testing time 2 (after the intervention study) for all tasks (three vocabulary knowledge tasks and two writing tasks). Because participants who could produce the 2K NGSL words before the intervention study started signifying that they had already acquired these words, they cannot be counted as their acquired words during the Quizlet study process. To compare the acquired words in writing topics, since I assigned two writing topics at two testing times, I mixed the two writing samples separately in testing times one and two. The acquired words in the writing samples are the results of excluding the same 2K NGSL words that appeared in testing time one and testing time 2 for two writing topics.

The results from Table 7.6 to Table 7.9 suggest that short-term studies can effectively improve vocabulary knowledge, particularly when using online flashcards. Additionally, the 'test task activation events', in which the chances occur that words are activated, significantly impact vocabulary knowledge elicitation (Fitzpatrick and Clenton, 2017). That study found that Lex30, which provides 30 cue words to access the lexical resources in participants' minds, resulted in the highest number of acquired words (n=278), followed by IELTS writing tasks, which offer topic carrier activation events (n=273), and G_Lex, which provides 24 sentence prompts to activate vocabulary knowledge (n=237). In contrast, the PVLT, a pre-determined vocabulary-filling task for each sentence, only elicited 22 acquired words, possibly due to its limited activation opportunities to access vocabulary knowledge. Overall, these findings suggest that using online flashcards can effectively improve vocabulary knowledge in a short amount of time, and vocabulary knowledge tasks with multiple chances

to access vocabulary knowledge, or the ‘capture zone’, elicit more acquired vocabulary knowledge from participants than vocabulary tasks with less ‘capture zone’.

Considering the findings shown in these tables, participants acquire vocabulary knowledge and can also use these words in their IELTS writing topics. I can thus infer that a lemma-based word list, such as the 2K NGSL word lists, is an appropriate resource to suit the level of the participants in chapter 6, and a bilingual version of word lists using flashcards could help participants learn words more easily. A short study period (12 weeks) significantly improves participants’ vocabulary and vocabulary knowledge in writing.

Table 7.6

Acquired Quizlet Words (2K NGSL Word List) Elicited by Lex30

Acquired NGSL words in Lex30 (type) (278)				
accident	cool	fuel	online	smoke
advertisement	copy	funny	opposite	snow
advice	crash	gas	pain	soft
advise	crowd	gift	pair	soldier
afford	cry	glad	partner	solve
angry	cup	glass	passenger	speed
announce	danger	global	path	split
appoint	dangerous	gold	peace	spring
appointment	debt	graduate	plane	stone
artist	defend	gun	plastic	storm
atmosphere	delay	handle	plate	suit
attitude	desk	hat	pleasure	sun
aware	destroy	heat	pool	swim
bag	device	heavy	pop	switch
ball	diet	hire	print	taxi
battle	dinner	hurt	quick	tea
beach	disappear	ice	quiet	tear
beat	disappoint	ideal	rain	telephone
beauty	dish	ill	rare	temperature
bike	disk	illness	regulation	text
bind	driver	illustrate	relax	thin
bird	dry	inform	rely	ticket
blow	ear	instrument	rent	tip
boat	earth	insurance	repair	tire

bomb	east	island	reserve	tone
bond	educate	joke	reveal	tool
bone	egg	journey	reward	topic
boring	elect	jump	rich	traffic
borrow	electronic	kick	ride	transport
boss	email	knock	river	troop
bottle	emotion	lake	roll	trust
brain	employ	leg	root	truth
breakfast	employment	library	route	vast
bright	empty	license	sad	vehicle
broad	enemy	literature	safe	video
budget	engine	load	safety	vision
burn	equipment	loan	salary	volume
camp	examination	locate	scale	volunteer
carefully	excite	location	schedule	wage
cash	exciting	lock	secret	wake
cat	expensive	mail	seed	warm
chart	expression	map	severe	wash
cheap	familiar	math	sheet	wave
classic	fan	meat	ship	weak
coach	farm	mechanism	shock	weapon
coast	fee	medicine	shoe	weather
coat	fellow	mistake	shoot	wheel
coffee	flat	mobile	shout	wind
colleague	flight	motion	shut	wine
comfort	flow	mountain	sick	winter
comfortable	flower	mouth	sight	wood
concert	forest	musical	signal	writer
conclude	frame	north	skin	yellow
conflict	freeze	notion	sky	youth
consequence	fresh	novel	slow	
cook	friendly	observation	slowly	

Table 7.7

Acquired Quizlet Words (2K NGSL Word List) Elicited by G_Lex

Acquired NGSL words in G_Lex (type) (237)				
accident	confirm	fashion	mistake	sentence
achievement	conflict	fat	monitor	severe
actor	confuse	flower	mountain	ship
admit	consequence	football	museum	shirt
advice	consideration	forest	online	shock
advise	construct	friendly	oppose	shoe
aid	content	fruit	ordinary	sick

amaze	context	funny	outcome	signal
angry	convince	gap	package	sky
announce	cook	gift	participant	smell
appearance	cool	glad	participate	smoke
appointment	copy	glass	partner	spring
appreciate	corporation	grade	perfect	status
approve	count	grateful	photo	steal
audience	crop	guard	photograph	stone
award	cry	guest	pleasure	strange
background	dangerous	habit	pool	succeed
bag	debt	handle	powerful	suit
ball	declare	hang	practical	suitable
ban	defeat	hate	preserve	sun
beach	defend	healthy	prevent	sweet
bedroom	delay	hide	print	swim
bike	delight	household	progress	talent
bird	deliver	hurt	promote	tall
birth	description	ignore	protest	tennis
bore	desk	illness	proud	terrible
boring	destroy	import	quiet	text
boss	diet	impression	rank	theater
bother	dinner	improvement	reaction	ticket
brain	disappear	incident	reasonable	tie
breakfast	disappoint	inform	reform	tire
burn	drama	introduction	reject	title
busy	egg	invite	relax	tool
calm	email	kiss	remark	topic
camera	emails	lake	remind	tough
careful	emotion	latter	rent	trust
cat	encounter	leg	repair	valuable
chair	enemy	library	reply	victory
cheap	escape	locate	request	video
climb	excellent	lovely	rich	wage
coach	exchange	luck	river	weak
coffee	excite	lucky	sad	wedding
comfortable	exciting	lunch	safe	wind
command	expensive	mail	salary	wine
commit	explanation	marriage	satisfy	wonderful
communicate	failure	math	scene	
complain	famous	meal	schedule	
conduct	fan	minister	select	

Table 7.8

Acquired Quizlet Words (2K NGSL Word List) Elicited by the PVLT

Acquired NGSL words in the PVLT (22)				
aware	democracy	justice	sequence	vision
climb	draft	nurse	sex	wage
connect	ensure	participate	surround	
copy	examine	phase	tip	
crisis	intelligence	root	usual	

Table 7.9

Acquired Quizlet Words (2K NGSL Word List) Elicited in Writing Tasks

Acquired NGSL words in writing tasks (type) (273)				
accident	confirm	fault	origin	significantly
achievement	confuse	fee	originally	slowly
acquire	connect	formal	otherwise	solve
active	consideration	frequently	ought	somewhat
actor	construct	funny	ourselves	spirit
actual	content	furthermore	overall	spot
adopt	context	gap	pain	spread
advertise	cool	gas	participant	strange
afraid	copy	global	participate	strength
amaze	corner	graduate	partly	stress
angry	corporation	habit	partner	stretch
anybody	correct	hardly	peak	strongly
anywhere	count	healthy	personality	succeed
appearance	crowd	heavy	physical	sudden
appoint	cultural	helpful	pleasure	suit
appropriate	cup	hide	pool	suitable
association	currently	highly	pop	surely
atmosphere	cycle	historical	practical	surroundings
attitude	dangerous	hurt	predict	survey
attract	decline	illness	prevent	swim
attractive	decrease	implement	progress	swing
audience	definitely	imply	promote	talent
aware	device	importance	radio	taxi
background	directly	impossible	rare	telephone
belief	disappear	impression	reaction	tennis
belong	distance	inform	reality	text

besides	distinguish	injury	recommend	threaten
bike	drama	insist	reduction	tie
blood	duty	intend	relax	tip
bond	earth	interaction	relevant	tire
bore	educate	invite	rely	tone
boring	effective	jump	remark	tool
bright	efficient	launch	rent	topic
busy	electronic	leg	reply	transport
camera	element	length	resolve	trend
cancer	email	lesson	respond	truly
careful	emails	mail	reveal	trust
carefully	emotion	mainly	rich	truth
channel	emotional	medical	ride	unable
childhood	enable	medicine	river	unless
chip	encounter	mental	rural	urge
circumstance	engage	minimum	sad	usual
citizen	equipment	mistake	safe	valuable
climb	everywhere	mobile	satisfy	vary
coach	examination	moreover	scale	video
colleague	exchange	mostly	schedule	wake
combination	excite	mouth	secondly	warm
combine	exciting	native	select	wash
comfortable	expand	neck	self	wave
commercial	expense	negative	senior	weak
commit	explanation	neighborhood	sentence	weather
communicate	expression	nobody	shoe	winter
concentrate	famous	normally	sick	wonderful
conduct	fashion	online	sight	
confidence	fat	ordinary	signal	

Furthermore, to validate how many words participants acquired individually, I chose four representative participants to present their gained vocabulary knowledge from the 2K NGSL word list. Table 7.10 shows the results of the acquired words from the 2K NGSL word list for the four randomly selected participants. In the Lex30 elicitation task, the first participant (s1) produced 23 words that were gained from the NGSL word lists. Out of these, s1 could use 16 words from the NGSL in their written production. The G_Lex elicitation task could elicit 11 words from s1, while the PVLT elicitation task elicited nine words. In the

Lex30 elicitation task, the second participant (s2) produced 23 NGSL words, out of which they could use 16 words in their written production. S2 produced two words in the PVLT elicitation task. The third participant (s3) demonstrated a strong performance in the Lex30 task by producing 36 NGSL words, 17 in the PVLT elicitation task, and 15 in the G_Lex elicitation task. However, s3 could only use 11 NGSL words in their written production. The fourth participant (s4) produced 30 NGSL words in the Lex30 elicitation task, along with 13 words in the PVLT task and 11 in the G_Lex task. S4 demonstrated the ability to use 14 NGSL words in written production. Overall, the results show the participants had varying proficiency levels in using NGSL words in their written production, and Lex30 proved to be the most successful task in eliciting NGSL words, rather than G_Lex or the PVLT.

Table 7.10

Examples of Acquired Quizlet Words (2K NGSL Word List)

Lex30 (s1)	burn classic defend dish gift glass ice mouth novel ride sad salary shout sight sky snow swim tear thin wage weather winter writer (23)
G_Lex (s1)	content diet disappoint drama fruit inform invite mail perfect quiet relax (11)
PVLT (s1)	aware climb copy justice nurse participate sequence usual wage (9)
Writing (s1)	commit communicate directly fee furthermore mainly mobile senior shoe sick spread tennis tire tool usual video (16)
Lex30 (s2)	boat bone bright elect gas glad graduate ice locate mouth novel root shoe sky tear temperature video volume wake warm wave wheel wind (23)
G_Lex (s2)	award burn defeat destroy exciting famous forest fruit illness mail photo remark reply satisfy swim tennis (16)
PVLT (s2)	root usual (2)
Writing (s2)	aware busy communicate disappear emotion examination expression fat habit mobile pool suit swim tool (14)
Lex30 (s3)	appointment bag beach bind burn cat coast cook debt dish earth enemy expensive familiar flight frame fuel gas gun heat heavy ice instrument loan mountain reserve shoot shout sight swim troop truth vision weapon wind winter (36)
G_Lex (s3)	amaze appointment beach boring climb debt delight destroy emotion fruit glad mistake print shock swim (15)
PVLT (s3)	aware climb copy crisis democracy draft ensure justice nurse participate phase root sex tip usual vision wage (17)
Writing (s3)	accident communicate construct graduate mobile native origin participate rent strange winter (11)
Lex30 (s4)	advice bag beach borrow bottle camp cat cook crash cup disappoint disk gas glass heat mistake online pool rent river shock snow swim text tire weapon wind wine winter wood (30)
G_Lex (s4)	bike disappoint exciting fruit funny handle mail pool promote shock theater (11)
PVLT (s4)	aware climb copy democracy draft ensure justice nurse participate sex surround usual wage (13)
Writing (s4)	belief communicate emotion expression impression mobile negative reduction sentence somewhat surely surroundings text topic (14)

The discussion sections mentioned above have addressed five main questions related to the four experimental chapters. First, I have examined the discrepancies in assessing vocabulary used in written production by vocabulary knowledge measures. Second, I have explored the selection of word counting units in vocabulary task responses and lexical diversity measures. Third, I have analysed how vocabulary knowledge tasks should be scored to assess vocabulary knowledge accurately. Fourth, I have investigated the relations among lexical diversity measures and which measures perform better in my experimental chapters. Fifth, I have considered the acquisition of words over a short period of study.

7.8 Limitations of the Study

Although the current dissertation provides valuable insights into the relations between vocabulary knowledge measures, lexical diversity measures, IELTS writing levels, and vocabulary knowledge development, several limitations need to be acknowledged. These include (1) the lack of comparison between the CEFR levels and IELTS writing scores; (2) the fixed text length used for lexical diversity scores; (3) using only the lemma/flemma word unit in a single chapter; (4) an insufficient number of high proficiency participants; and (5) the intervention chapter being limited by the single language background (L1 Japanese) participants and the length of the intervention, which could potentially impact the generalisability of the findings.

The study has these potential limitations, which I will explain further. First, there is a lack of comparison between CEFR levels and IELTS. Even though I have reported general

CEFR levels for the participants in my experimental chapters, their CEFR levels were subjectively judged by their English language instructors. I have not accessed their actual English language standardised test scores (e.g., IELTS or TOEIC). The CEFR levels contain four aspects: listening, speaking, reading, and writing. The current dissertation mainly focuses on participants' written production, and if I used their actual CEFR levels, it may diminish some potential factors in assessing their writing level. Chapter 5 uses qualified IELTS raters to judge the IELTS writing scores. A study designed to compare CEFR levels and IELTS writing scores would benefit the research community.

Second, one limitation of the current study is the text length. The essay length for the current study is generally between 200-400 English words. The current study analysed LD scores based on the representative sampling of the middle 200 English words, as Treffers-Daller et al. (2018) conducted, since text length is a common problem within lexical diversity measures. The more recently created lexical diversity measures claim not to be affected by text length, but different text lengths can still influence the lexical diversity scores. To ensure the calculation results in my current dissertation excluded the influence of the text length problem, I chose a consistent text length (the middle 200 English words) for all writing samples. Despite extracting the middle 200 English words, the LD scores inevitably varied for every 200 words of an essay over 200 words. To explain, since LD scores are determined by adjusting the number of types and tokens, any difference in the proportion of types used for different parts of an essay can cause differences in the LD score, so a random 200-word sample selection from an essay does not reflect an equivalent proportion of types from a specific participant. As a result, ensuring an equal proportion of types in 200-word samples

during the writing process is likely impossible. For example, if I were to divide a 2000-word essay into ten sections (e.g., 200 words each), participants would not use the same types throughout each section. Despite such a concern, future studies should investigate different text lengths or include all words within the data with different window sizes to explore how different text lengths would influence lexical diversity scores, or the role that window size plays in affecting lexical diversity scores, as explored in Treffers-Daller et al. (2022).

Third, it should be noted that the current dissertation only uses lemma/flemma word counting units and a lemma-based word list for the intervention study. As mentioned above, the choosing of appropriate word units relates to participants' proficiency levels. Recent studies (e.g., McLean, 2018; Brown et al., 2020) indicated that the flemma/lemma is an appropriate word unit for L1 Japanese participants, and the participants in the current dissertation are mostly from Japan. I assume they lack the language ability to fully understand derived forms of words. Nevertheless, the current dissertation has not distinguished between participants' knowledge of word families and their overall proficiency. Although certain studies (Milton & Alexiou, 2009; Nation, n.d.) have shown that language learners' word family size is related to both their vocabulary size and CEFR levels, they have also suggested different vocabulary sizes among participants with similar CEFR levels. Future studies can investigate the relations between different CEFR levels, vocabulary sizes, and participants' word family levels. The Word Part Levels Test (WPLT; Sasao & Webb, 2017) may also offer a solution to evaluate participants' word family knowledge.

Fourth, a potential limitation concerns the number of participants with high proficiency levels. The participants' IELTS writing scores in chapter 5 ranged from 5.5 to

6.5. The number of participants whose IELTS writing scores are equal to and over 6.5 is 21, compared with those whose IELTS writing scores fall at 6.0 (n=49) and 5.5 (n=28). Including more participants with higher IELTS writing levels might help to maintain a balanced data distribution for the number of participants in each group. However, a large proportion of participants in my dissertation come from Japan, and most of them are undergraduates; it is uncommon to find numerous such participants with a high English language proficiency level.

Fifth, the study reported in the intervention study (chapter 6) shows no significant correlations between productive vocabulary knowledge task scores and lexical diversity measure scores in the post-test data. A reason for this lack of correlations may be the length of the intervention and the waning motivation of participants towards word list studies. The study reported in chapter 6 is potentially limited by the single language background (L1 Japanese) participants and the length of the intervention. First, the study reported in chapter 6 only includes L1 Japanese participants who are undergraduates at a university in Japan. Their English language proficiency levels are B1. The number of participants was 51. I was not able to include participants from other language backgrounds with different proficiency levels. The findings might differ if chapter 6 included participants from diverse language backgrounds with different language proficiency levels. These factors might be worthy of follow-up studies. Additionally, the total period of intervention for the intervention study was 12 weeks, as it was in the intervention study conducted by Cobb and Horst (2001), but some studies have also examined vocabulary knowledge growth in language learners throughout a year-long study period (e.g., Daller et al., 2013; Fitzpatrick, 2012). Even though the

participants were required to learn the words using Quizlet during out-of-class time, it remains unknown whether they did or not.

7.9 Implications for Pedagogy and Assessment

Despite the current dissertation having some limitations, understanding the implications of the findings can provide valuable insights into the practical applications from two broad viewpoints. First, the findings provide essential implications for pedagogical, especially the L2 writing classes. Second, the findings also provide significant implications for vocabulary knowledge assessment.

First, the findings are important for L2 writing classes. The findings discussed in section 7.3.2 showed significant correlations between vocabulary knowledge scores and lexical diversity scores in IELTS writing production. Participants with higher vocabulary knowledge scores would also gain higher lexical diversity scores in their writing. Section 7.3.3 discussed how vocabulary knowledge tasks with more embedded features showed closer relationships lexical diversity scores. Furthermore, section 7.3.4 discussed how vocabulary knowledge scores could differentiate between three IELTS writing scores, and participants with higher IELTS writing scores demonstrated closer relationships between vocabulary knowledge scores and lexical diversity scores than did lower-level participants. The findings discussed in sections 7.3.2, 7.3.3, and 7.3.4 reveal positive linear relationships among vocabulary knowledge scores, lexical diversity scores, and IELTS writing scores. This indicates that increasing language learners' vocabulary knowledge appears to be an effective way to improve both their lexical diversity scores in written production and their IELTS

writing scores. Language instructors can focus on teaching more words beyond the 2K level and providing more opportunities for students to use vocabulary knowledge under various contexts. Moreover, the findings suggested that language instructors could consider using productive vocabulary knowledge tasks (Lex30, G_Lex, and the PVLT) to evaluate learners' writing levels. The findings may give language educators a better understanding of the status and development of a language learner's vocabulary knowledge at various writing levels.

Further, the findings highlight the importance of productive vocabulary tasks in assessing participants' proficiency levels in writing. A writing teacher might be interested in knowing the score that shows the student's proficiency level of, for instance, IELTS 5.0, IELTS 6.0, or IELTS 7.0. The G_Lex test, which was found to be the most accurate predictor of productive vocabulary knowledge for participants among the four experimental chapters, contains a maximum of 120 words of all responses in a single G_Lex task. It is pertinent to inquire whether a rough estimate of the threshold score can be established based on future studies. For example, would correctly filling out 55 out of the potential 120 spaces on the G_Lex indicate a proficiency level of IELTS 5.0 or 5.5? This information would be valuable for language instructors and institutions in determining the appropriate ability classes for students.

In addition, emphasising the teaching or training of word part levels knowledge can significantly improve language learners' ability to comprehend and use vocabulary. The current dissertation uses a lemma or flemma word counting unit in consideration of a recent article by McLean (2018), as discussed in section 7.4. His study shows that L1 Japanese participants had insufficient vocabulary knowledge in understanding word families,

especially with the derived forms. In teaching practice, even though language learners have been taught these words, teachers rarely stress word part knowledge in teaching activities. Dang (2021), for instance, suggests that there is a lack of training in the knowledge of word parts and that through learning frequently derived word knowledge items, learners can understand the importance of morphological knowledge and increase their understanding of vocabulary. If language instructors can stress the significance of teaching word part knowledge, it may help learners learn words and improve their language skills. In support, Webb (2021) suggested that 'presenting headwords together with their inflections and derivations may provide a shortcut to lexical development' (p. 942). Such a claim also highlights that acquiring morphological knowledge can improve vocabulary knowledge learning in practice.

In addition, conducting a longitudinal study on utilising word lists to improve writing scores over a short study period would be worthwhile. The findings discussed in section 7.7 regarding the vocabulary acquired from the NGSL showed that participants could produce the 2K NGSL words in three productive vocabulary tasks as well in their IELTS writing production. Through a pre- and post-test design, only G_Lex could track vocabulary knowledge growth, whereas all lexical diversity measures could track vocabulary knowledge growth in different IELTS writing topics. These findings have practical implications in pointing out the effectiveness of using word lists to improve writing proficiency. Language programs that aim to improve learners' writing performance can use this method in their curriculum development. The study also highlights the importance of using longitudinal research methods to evaluate word lists in writing, which can inform instructional practices.

Having outlined the implications for L2 writing classes, I shift the focus in the following paragraphs to the implications of the current dissertation concerning vocabulary knowledge assessment.

Second, the findings also provide significant implications for vocabulary knowledge assessment. The findings in sections 7.3.2, 7.3.3, and 7.3.4 show that vocabulary knowledge tasks (Lex30, G_Lex, and the PVLT) demonstrated discrepancies in predicting lexical diversity scores and differentiating participants at different IELTS writing scores. Lex30 and the PVLT task scores show closer relationships with participants getting the highest IELTS writing scores, whereas the G_Lex scores show greater ability with lexical diversity scores in chapter 4. Only the G_Lex scores could track vocabulary knowledge development for participants over a short study period.

The agreement among vocabulary researchers is that higher proficiency participants demonstrate greater knowledge in their vocabulary: vocabulary knowledge in use (lexical diversity scores), and word part knowledge (in both inflected and derived forms) than lower-level participants. The current dissertation uses lemma and flemma as word counting units for vocabulary knowledge tasks and writing samples. Previous studies (Edmonds et al., 2022; Meara & Fitzpatrick, 2000; Fitzpatrick & Clenton, 2017) used lemma as a word unit for Lex30 and G_Lex, and their lemma standards equate to the level 2 and part of the level 3 word family level created by Bauer and Nation (1993). Furthermore, vocabulary tools can also play an effective role in eliciting vocabulary acquired from word lists. The findings discussed in section 7.7 indicated that Lex30 elicits the largest amount of vocabulary (278) acquired from the NGSL word lists, followed by writing samples (273), and then G_Lex

(237). The PVLT required participants to fill in one single, pre-determined word for each sentence and could elicit 22 words acquired from the NGSL word list. These findings indicate that Lex30 and G_Lex could be used as tools to track vocabulary knowledge improvement.

7.10 Future Research

Given the potential implications of vocabulary knowledge for pedagogy and assessment, further research is needed. As such, it is important to draw conclusions only from the existing evidence.

Currently, institutions rely on tests of general language proficiency, such as TOEFL and IELTS, to divide students into ability classes, which may not be very precise or skill specific. Administering specific tests for each skill, such as G_Lex for writing, would more effectively assess participants' proficiency levels. While this topic requires further empirical work to identify specific threshold scores for IELTS levels, it is worth validating in future research.

However, we also need further research on how morphological knowledge influences vocabulary learning. Even when both language instructors and language learners pinpoint the importance of learning word part knowledge, questions still remain regarding the word part knowledge that should be learned. In response, Nation and Bauer (2023) suggested that learning affixes was related to learners' gain in vocabulary size: they claim that level 3 affixes should be studied for learners who know the first 1,000 flemmas, level 4 affixes should be studied for learners who have acquired the first 2,000 and 3,000 words, and level 5

affixes and beyond need to be studied for learners who know the first 4,000 words. Nation and Bauer's (2023) claim offers a solution in teaching practice that language instructors should also consider learner vocabulary size when they present the knowledge of word families in their teaching activities. However, further empirical research is needed to substantiate claims about vocabulary size and word family levels.

Considering that the PVLT involves potential receptive vocabulary knowledge features, the current dissertation recommends that future research instead use Lex30 and G_Lex as practical tools to assess productive vocabulary knowledge. Lex30 presents better performance with high IELTS writing score participants than G_Lex, while G_Lex shows better performance than Lex30 with relatively lower proficiency level participants. Future studies can employ multiple vocabulary knowledge tasks in assessing the relationship between vocabulary knowledge and IELTS writing. These suggestions may only be suitable for writing, and I am uncertain of their practical implications for speaking, which is another important productive language skill.

In addition, to establish a standard word counting unit for the responses in vocabulary tasks for Lex30 and G_Lex, future empirical research can explore the relations between proficiency, word family levels, and vocabulary size. In addition, future research can validate Lex30 and G_Lex tools with different word lists and more participants of different language backgrounds or language skills.

Moreover, follow-up studies should include participants from diverse language backgrounds with different language proficiency levels to investigate vocabulary knowledge growth. Such studies would increase the generalizability of the findings. Follow-up studies

can increase our understanding of vocabulary knowledge growth, lexical diversity measures, and vocabulary knowledge measures. Including participants from different language backgrounds can enhance our understanding of whether lexical diversity and vocabulary knowledge measures are more effective in discriminating between participants from specific language backgrounds. By including participants with different language proficiency levels, we can examine how vocabulary knowledge development varies with proficiency. This may shed light on the relationship between vocabulary knowledge growth and overall language proficiency.

7.11 Summary of Findings

This research is the first to implement a multi-dimensional approach to vocabulary scores when predicting LD scores and differentiate between writing skills. In addition, it is the first study to investigate vocabulary knowledge development over a short time study period by vocabulary knowledge scores and LD scores. The discoveries have significant implications for pedagogy, language education, and vocabulary knowledge assessment. The current dissertation has evaluated productive vocabulary knowledge using writing scores and highlights the importance of adopting a multi-task approach to evaluating productive vocabulary knowledge in assessing writing proficiency. The dissertation also assessed vocabulary knowledge development using multiple vocabulary knowledge measures and LD measures to address the lack of longitudinal studies in vocabulary knowledge assessment. The positive results showed participants could acquire the NGSL words and then produce these words in both vocabulary knowledge tasks and IELTS writing.

Considering the discussions in this chapter, I will conclude this chapter by restating the main findings:

- Vocabulary knowledge tasks showed discrepancies in assessing vocabulary knowledge in use.
    - There were no significant correlations between vocabulary knowledge and lexical diversity measures for pre-intermediate participants (CEFR=A2). Low proficiency level participants appeared to lack the ability to put their vocabulary knowledge into writing use.
    - Higher proficiency participants showed closer relationships between their vocabulary knowledge scores and lexical diversity scores compared to lower proficiency participants.
    - The PVLT scores yielded the strongest and the most significant correlations with the LD measures compared to Lex30 scores and G_Lex task scores for intermediate to high proficiency level participants.
    - Vocabulary knowledge task scores (G_Lex and the PVLT) with more embeddedness features, i.e., involving more context, explained the greater variance in lexical diversity measures (with TTR index) than Lex30. G_Lex scores explained the highest percentage of the variance in TTR scores in chapter 4 for L1 Japanese participants whose levels ranged from intermediate to advanced. The PVLT scores explained the highest percentage of the variance in TTR scores in chapter 5 for L1 Japanese and L1 French participants whose proficiency ranged from intermediate to high.

    - The PVLT scores showed the highest R^2 values in distinguishing writing levels (judged by qualified IELTS raters), followed by Lex30 and G_Lex task scores.
    - Considering the PVLT task of obtaining receptive vocabulary knowledge features (Edmonds et al., 2022), the current dissertation recommends both G_Lex and Lex30 as better productive vocabulary knowledge predictors in predicting writing proficiency (lexical diversity scores).
- Traditional lexical diversity measures such as Types, TTR, Root_TTR, Log_TTR, MAAS, and MSTTR showed closer relationships with productive vocabulary measures. However, more recently devised lexical diversity measures such as MTLD_W, MTLD, and MATTR showed a greater effect size in tracking vocabulary knowledge development than those previously deployed measures.
- Keeping word counting units consistent for both vocabulary knowledge measures and lexical diversity measures was essential in the study. The current dissertation used lemma as a word counting unit in chapter 3 and flemma as a word counting unit in chapters 4, 5, and 6 for responses from vocabulary knowledge tasks and writing topics.
- Using lemma/flemma as a word counting unit for vocabulary knowledge measures showed concurrent validity with published studies.
- The G_Lex task indicated greater power in tracking vocabulary knowledge improvement than either the PVLT or Lex30.
- All lexical diversity measures tracked vocabulary knowledge growth in writing, and MTLD_W showed the largest effect size among the 11 lexical diversity indices.

- Short-time intervention study using online flashcards was an effective way to improve vocabulary knowledge and vocabulary knowledge use in writing.
- Lex30 elicited the highest number (278) of acquired vocabulary knowledge items over a short study period, followed by G_Lex (237) and then the PVLT (22).
- Participants utilised a significant number (273) of the vocabulary items acquired from the New General Service List (NGSL) in their written production.

Chapter 8: Conclusion

This dissertation is the first to implement a multidimensional approach to vocabulary task scores in predicting lexical diversity scores and differentiating between writing scores. It is also the first study to investigate vocabulary knowledge development with vocabulary and lexical diversity scores over a short study period. The findings provide significant implications for pedagogy, language education, and vocabulary knowledge assessment. The dissertation highlights the importance of adopting a multi-task approach to evaluating productive vocabulary knowledge in writing proficiency. Because of the lack of longitudinal studies in vocabulary knowledge assessment, it also assesses vocabulary knowledge development using a battery of vocabulary and lexical diversity measures.

This dissertation has explored how vocabulary knowledge measures can manifest their predictive capability in evaluating vocabulary in use in written activities for participants of different proficiency levels. I have investigated how vocabulary knowledge measures can distinguish between different IELTS writing levels. I have also examined how vocabulary knowledge measures and lexical diversity measures can track vocabulary knowledge development over a short study time. To better investigate these questions, I employed multiple vocabulary knowledge measures because of their multidimensionality. In my discussion chapter, I address the nature of these various dimensions in greater detail.

A further unique element of this dissertation is the evaluation of different word type counting. In a partial replication of the Treffers-Daller et al.'s (2018) study, I adhered to the lemma as the word unit both for responses from the vocabulary tasks and for writing samples in my first experimental chapter (chapter 3). Then, on the basis that Treffers-Daller et al. did

not explore the use of flemma as a word unit, I have used it as a word counting unit for both the vocabulary tasks and written production in my experimental chapters 4–6.

Based on my experimental chapters' findings, the discussion chapter (chapter 7) extrapolated and examined six essential issues. I highlighted the main findings concerning the current dissertation. First, vocabulary knowledge measures show discrepancies in accessing vocabulary use in written production, and this may be caused by two main factors: vocabulary knowledge tasks differ in their characteristics, and they differ in the degree of their embeddedness. Second, the vocabulary knowledge measures utilised in the current dissertation can distinguish between participants with different IELTS writing scores. Third, G_Lex shows better performance in tracking vocabulary knowledge development than either the PVLT or Lex30. Fourth, online flashcard learning with a lemma-based word list is shown to be an effective way to improve word learning. Fifth, I have also highlighted that lexical diversity measures can present different predictive power depending on the research question: traditional lexical diversity measures show closer relationships with vocabulary task scores in writing use; however, more recently created lexical diversity measures show better performance in tracking vocabulary knowledge development. Sixth, the correlation results among vocabulary task scores and among lexical diversity measure scores conducted in the current dissertation show concurrent validity with published studies.

I highlighted the essential implications from the current dissertation from two broad perspectives in my discussion chapter. First, the findings provide important implications for L2 writing classes, with the suggestion that increasing language learners' vocabulary knowledge appears to be an effective way to improve both their lexical diversity scores in

written production and their IELTS writing scores. Thus, language instructors can focus on teaching more words beyond the 2K level and allow more opportunities for students to use their vocabulary knowledge in various contexts. Second, the findings emphasised the importance of using low-stakes productive vocabulary knowledge tasks (Lex30, G_Lex, and the PVLT) to predict and distinguish writing levels. These vocabulary task scores can explain different percentages of the variance in lexical diversity scores. Third, the findings also provide implications for vocabulary knowledge assessment in that teachers should employ multiple vocabulary tasks in assessing written production. The results in this dissertation suggest that the different vocabulary tasks have different strengths of significant correlations for participants with different proficiencies.

References

Airey, J. (2015). From stimulated recall to disciplinary literacy: Summarizing ten years of research into teaching and learning in English. In S. Dimova, A. K. Hultgren, & C. Jensen (Eds.), *English-medium Instruction in European Higher Education*, (pp. 157–176). De Gruyter Mouton. https://doi.org/10.1515/9781614515272

Alqahtani, M. (2015). The importance of vocabulary in language learning and how to be taught. *International Journal of Teaching and Education*, *3*(3), 21–34.

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. Guthrie (Ed.), *Comprehension and Teaching: Research Reviews*, (pp. 77–117). International Reading Association.

Anthony, L. (1999). Writing research article introductions in software engineering: How accurate is a standard model?. *IEEE Transactions on Professional Communication*, *42*(1), 38–46.

Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*, *30*(2), 141–161.

Anthony, L. (2022). AntWordProfiler (Version 2.0.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Arnaud, P. J. L. (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* (pp. 14–28). University of Essex.

Arnaud, P. J. L. (1992). Objective lexical and grammatical characteristics of L2 written compositions and the validity of separate-component tests. In P. J. L., Arnaud, & H.

Béjoint (Eds.), *Vocabulary and Applied Linguistics* (pp. 133–145). Palgrave Macmillan. https://doi.org/10.1007/978-1-349-12396-4_13

Baba, K. (2002). Test review: Lex30. *Language Testing Update*, *32*, 68–71.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.

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

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing*, *16*(2), 131–162. https://doi.org/10.1177/026553229901600202

BNC Consortium. (2007). British national corpus. *Oxford Text Archive Core Collection*.

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

British Council. (2022, August 4). *IELTS TASK 2 Writing band descriptors (public version)*. The British Council, IDP. https://www.ielts.org/-/media/pdfs/writing-band-descriptors-task-2.ashx

Brown, D., Stoeckel, T., McLean, S., & Stewart, J. (2020). The most appropriate lexical unit for L2 vocabulary research and pedagogy: A brief review of the evidence. *Applied Linguistics*, *43*(3), 596–602. https://doi.org/10.1093/applin/amaa061

Browne, C. (2014). A new general service list: The better mousetrap we've been looking for? *Vocabulary Learning and Instruction*, *3*(2), 1–10. http://dx.doi.org/10.7820/vli.v03.2.browne

Browne, C. (2021). The NGSL project: Building wordlists and resources to help EFL learners (and teachers) to succeed. In E. Forsythe (Ed.), *Teaching with Technology 2020 Selected papers from the JALTCALL2020 Conference* (pp. 1–18).

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Carroll, J. B. (1938). Diversity of vocabulary and the harmonic series law of word-frequency distribution. *Psychological Record*, *2*, 379–386.

Carroll, J. B. (1964). *Language and thought*. Prentice-Hall.

Catalán, R. M. J., & Llach, M. P. A. (2017). CLIL or time? Lexical profiles of CLIL and non-CLIL EFL learners. *System*, *66*, 87–99. https://doi.org/10.1016/j.system.2017.03.016

Catala, R. J., & Espinosa, S. M. (2005). Using Lex30 to measure the L2 productive vocabulary of Spanish primary learners of EFL. *Vigo International Journal of Applied Linguistics*, 2, 27–44.

Caton, T. H. (2018). Short-Term Study Abroad and Lex30: a Replication Study. *Bulletin of Nakamura Gakuen University and Nakamura Gakuen Junior College*, 50, 123–131.

Chang, A. C. S. (2007). The impact of vocabulary preparation on L2 listening comprehension, confidence and strategy use. *System*, *35*(4), 534–550. https://doi.org/10.1016/j.system.2007.06.003

Chapelle, C. A. (2006). L2 vocabulary acquisition theory: The role of inference, dependability and generalizability in assessment. In M. Chalhoub-Deville, C. A. Chapelle, & P. A. Duff (Eds.), *Inference and Generalizability in Applied Linguistics* (pp. 47–64). John Benjamins Publishing Company. http://dx.doi.org/10.1075/lllt.12.05cha

Clenton, J. (2005). Why Lex30 may not be an improved method of assessing productive vocabulary in an L2. *Studies in Language and Culture*, 31, 47–59.

Clenton, J. (2010). *Investigating the construct of productive vocabulary knowledge with Lex30* (Publication No. 10797989). [Doctoral dissertation, Swansea University]. ProQuest Dissertations. https://www.proquest.com/dissertations-theses/investigating-construct-productive-vocabulary/docview/2008472046/se-2

Clenton, J., de Jong, N., Clingwall, D., & Fraser, S. (2020). Investigating the extent to which vocabulary knowledge and skills can predict aspects of fluency for a small group of pre-intermediate Japanese L1 users of English (L2). In J. Clenton, & P. Booth (Eds.), *Vocabulary and the Four Skills Pedagogy, Practice, and Implications for Teaching Vocabulary* (pp. 126–145). Routledge.

Cobb, T., & Horst, M. (2001). Growing academic vocabulary with a collaborative on-line database. In B. Morrison, D. Gardner, K. Keobke, & M. Spratt (Eds.), *ELT Perspectives on Information Technology & Multimedia: Selected Papers from the ITMELT 2001 Conference 1st & 2nd* (pp. 189–226). Hong Kong Polytechnic University.

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

Cohen, A. D. (1989). Attrition in the productive lexicon of two portuguese third language speakers. *Studies in Second Language Acquisition*, *11*(2), 135–149. http://doi.org/10.1017/S0272263100000577

Corson, P. (1995). *Using English words*. Springer Science & Business Media.

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

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94–100. https://doi.org/10.1080/09296171003643098

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

Coxhead, A. (2012). Academic vocabulary, writing and English for academic purposes: Perspectives from second language learners. *RELC Journal*, *43*(1), 137–145. https://doi.org/10.1177/0033688212439323

Daller, H. (1999). Migration und Mehrsprachigkeit. *Der Sprachstand türkischer Rück kehrer aus Deutschland, Frankfurt am Main: Lang*.

Daller, M. H., & Phelan, D. (2013). Predicting international student study success. *Applied Linguistics Review*, *4*(1), 173–193. https://doi.org/10.1515/applirev-2013-0008

Daller, M., 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. 187–217). John Benjamins Publishing Company.

Daller, H., & Xue, H. (2007). Lexical richness and the oral proficiency of Chinese EFL students. In D. Helmut, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. 150–164). Cambridge University Press.

Dalton-Puffer, C. (2011). Content-and-Language Integrated Learning: From Practice to Principles? *Annual Review of Applied Linguistics*, *31*, 182–204. http://doi.org/10.1017/S0267190511000092

Dang, T. N. Y. (2020). The potential for learning specialized vocabulary of university lectures and seminars through watching discipline-related TV programs: insights from medical corpora. *TESOL Quarterly*, *54*(2), 436–459. https://doi.org/10.1002/tesq.552

Dang, T. N. Y. (2021). Selecting lexical units in wordlists for EFL learners. *Studies in Second Language Acquisition*, *43*(5), 954–957. https://doi.org/10.1017/S0272263121000681

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, *25*(4), 447–464. https://doi.org/10.1093/llc/fqq018

de Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, *36*(2), 223–243. https://doi.org/10.1017/S0142716413000210

de Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*(1), 5–34. https://doi.org/10.1017/S0272263111000489

Dickinson, D. K., & Tabors, P. O. (2001). *Beginning literacy with language: Young children learning at home and school*. Paul H Brookes Publishing.

Durán, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics*, *25*(2), 220–242. https://doi.org/10.1093/applin/25.2.220

Earls, C. W. (2016). *Evolving agendas in European English-medium higher education: Interculturality, multilingualism and language policy*. Springer.

Edmonds, A., Clenton, J., & Elmetaher, H. (2022). Exploring the construct validity of tests used to assess L2 productive vocabulary knowledge. *System*, *108*, 102855. https://doi.org/10.1016/j.system.2022.102855

Elgort, I. (2018). Technology-mediated second language vocabulary development: A review of trends in research methodology. *Calico Journal*, *35*(1), 1–29. https://www.jstor.org/stable/90016519

Engber, C. (1993). *The relationship of lexis to quality in L2 compositions* [Paper presentation]. TESOL'93, Atlanta, Georgia.

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

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

Fergadiotis, G., Wright, H. H., & West, T. M. (2013). Measuring lexical diversity in narrative discourse of people with aphasia. *American Journal of Speech-Language Pathology*, *22*(2), 397–408. https://doi.org/10.1044/1058-0360

Fergadiotis, G., Wright, H. H., & Green, S. B. (2015). Psychometric evaluation of lexical diversity indices: Assessing length effects. *Journal of Speech, Language, and Hearing Research*, *58*(3), 840–852. https://doi.org/10.1044/2015_JSLHR-L-14-0280

Fitzpatrick, T. (2006). Habits and rabbits: Word associations and the L2 lexicon. *EUROSLA Yearbook*, *6*(1), 121–145. https://doi.org/10.1075/eurosla.6.09fit

Fitzpatrick, T. (2007). Productive vocabulary tests and the search for concurrent validity. In D. Helmut, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. 116–133). Cambridge University Press. https://doi.org/10.1017/CBO9780511667268.009

Fitzpatrick, T. (2012). Tracking the changes: Vocabulary acquisition in the study abroad context. *The Language Learning Journal*, *40*(1), 81–98. https://doi.org/10.1080/09571736.2012.658227

Fitzpatrick, T., & Clenton, J. (2010). The challenge of validation: Assessing the performance of a test of productive vocabulary. *Language Testing*, *27*(4), 537–554. https://doi.org/10.1177/0265532209354771

Fitzpatrick, T., & Clenton, J. (2017). Making sense of learner performance on tests of productive vocabulary knowledge. *TESOL Quarterly*, *51*(4), 844–867. https://doi.org/10.1002/TESQ.356

Fitzpatrick, T., & Meara, P. (2004). Exploring the validity of a test of productive vocabulary. *Vigo International Journal of Applied Linguistics*, 1, 55–74.

Gablasova, D., & Brezina, V. (2021). Words that matter in L2 research and pedagogy: A corpus-linguistics perspective. *Studies in Second Language Acquisition*, *43*(5), 958–961. http://doi.org/10.1017/S027226312100070X

Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics*, *28*(2), 241–265. https://doi.org/10.1093/applin/amm010

González, R. A., & Píriz, A. M. P. (2016). Measuring the productive vocabulary of secondary school CLIL students: Is Lex30 a valid test for low-level school learners?. *Vigo International Journal of Applied Linguistics*, 13, 31–54.

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be?. *Applied Linguistics*, *11*(4), 341–363. https://doi.org/10.1093/applin/11.4.341

Grabe, W., & Stoller, F. L. (1997). Reading and vocabulary development in a second language. In J. Coady & T. Huckin (Eds.), *Second Language Vocabulary Acquisition* (pp. 98–122). Cambridge University Press.

Gu, P. Y. (2003). Vocabulary learning in a second language: Person, task, context and strategies. *TESL-EJ*, *7*(2), 1–25. http://teslej.org/wordpress/issues/volume7/ej26/ej26a4/

Guiraud, P. (1954). *Les caractères statistiques du vocabulaire: essai de méthodologie*. Presses Universitaires de France.

Hamrick, P. (2019). Adjusting Regression Models for Overfitting in Second Language Research. *Journal of Research Design & Statistics in Linguistics & Communication Science*, *5*, 107–122. https://doi.org/10.1558/jrds.38374

Heatley, A., & Nation, I. S. P. (1998). VocabProfile and range. *School of Linguistics and Applied Language Studies. Victoria University of Wellington, Wellington, New Zealand*.

Henriksen, B., & Danelund, L. (2015). Studies of Danish L2 learners' vocabulary knowledge and the lexical richness of their written production in English. In K. Doró, P. Pietilä & R. Pípalová (Eds.), *Lexical Issues in L2 Writing* (pp. 29–56). Cambridge Scholars Publishing.

Herdan, G. (1960). *Type-token mathematics: A textbook of mathematical linguistics*. Mouton and Co.

Huang, H. (2010). *How does second language vocabulary grow over time? A multi-methodological study of incremental vocabulary knowledge development* (Publication No. 3415908). [Doctoral dissertation, University of Hawaii at Manoa]. ProQuest Dissertations. https://www.proquest.com/dissertations-theses/how-does-second-language-vocabulary-grow-over/docview/734797551/se-2

Huhta, A., Alanen, R., Tarnanen, M., Martin, M., & Hirvelä, T. (2014). Assessing learners' writing skills in a SLA study: Validating the rating process across tasks, scales and

languages. *Language Testing*, *31*(3), 307–328. https://doi.org/10.1177/0265532214526176

Isbell, D. R., & Kremmel, B. (2020). Test review: Current options in at-home language proficiency tests for making high-stakes decisions. *Language Testing*, *37*(4), 600–619. https://doi.org/10.1177/0265532220943483

Iwaizumi, E., & Webb, S. (2021). To what extent does productive derivational knowledge of adult L1 speakers and L2 learners at two educational levels differ?. *TESOL Journal*, *12*(4), e640. https://doi.org/10.1002/tesj.640

Iwaizumi, E., & Webb, S. (2022). To what extent do learner-and word-related variables affect production of derivatives?. *Language Learning*. https://doi.org/10.1111/lang.12524

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19*(1), 57–84. https://doi.org/10.1191/0265532202lt220oa

Jarvis, S. (2013a). Defining and measuring lexical diversity. In S. Jarvis & M. Daller (Eds.), *Vocabulary Knowledge. Human Ratings and Automated Measures* (pp. 13–44). John Benjamins Publishing Company. https://doi.org/10.1075/sibil.47.03ch1

Jarvis, S. (2013b). Capturing the diversity in lexical diversity. *Language Learning*, *63*(s1), 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing*, *34*(4), 537–553. https://doi.org/10.1177/0265532217710632

Jarvis, S., & Hashimoto, B. J. (2021). How operationalizations of word types affect measures of lexical diversity. *International Journal of Learner Corpus Research*, *7*(1), 163–194. https://doi.org/10.1075/ijlcr.20004.jar

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, *64*(1), 160–212. https://doi.org/10.1111/lang.12034

Jiang, N. (2002). Form–meaning mapping in vocabulary acquisition in a second language. *Studies in Second Language Acquisition*, *24*(4), 617–637. https://doi.org/10.1017/S0272263102004047

Johnson, W. (1939). *Language and speech hygiene* (General Semantics Monographs, No.1). Institute of General Semantics.

Johnson, W. (1944). Studies in language behavior: A program of research. *Psychological Monographs*, *56*(2), 1–15.

Johnson, M. D., Acevedo, A., & Mercado, L. (2016). Vocabulary knowledge and vocabulary use in second language writing. *TESOL Journal*, *7*(3), 700–715. https://doi.org/10.1002/tesj.238

Kiliç, M. (2019). Vocabulary Knowledge as a Predictor of Performance in Writing and Speaking: A Case of Turkish EFL Learners. *PASAA: Journal of Language Teaching and Learning in Thailand*, *57*, 133–164.

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

Koizumi, R., & In'nami, Y. (2013). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *Journal of Language Teaching and Research*, *4*(5), 900–913. https://doi:10.4304/jltr.4.5.900-913

Kremmel, B. (2016). Word families and frequency bands in vocabulary tests: Challenging conventions. *TESOL Quarterly*, *50*(4), 976–987. https://doi.org/10.1002/tesq.329

Kremmel, B., & Pellicer-Sánchez, A. (2020). Measuring vocabulary development. In P. Winke & T. Brunfaut (Eds.), *The Routledge Handbook of Second Language Acquisition and Language Testing* (pp. 211–222). Routledge.

Kremmel, B., & Schmitt, N. (2016). Interpreting vocabulary test scores: What do various item formats tell us about learners' ability to employ words?. *Language Assessment Quarterly*, *13*(4), 377–392. https://doi.org/10.1080/15434303.2016.1237516

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Brown University Press.

Kyle, K. (2019). Measuring lexical richness. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 454–476). Routledge.

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, *18*(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

LaFlair, G. T., Egbert, J., & Plonsky, L. (2015). A practical guide to bootstrapping descriptive statistics, correlations, t tests, and ANOVAs. In L. Plonsky (Ed.), *Advancing Quantitative Methods in Second Language Research*, (pp. 46–77). Routledge.

Larson-Hall, J. (2015). *A guide to doing statistics in second language research using SPSS and R*. Routledge.

Larson-Hall, J., & Herrington, R. (2010). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics*, *31*(3), 368–390. https://doi.org/10.1093/applin/amp038

Laufer, B. (1991). The development of L2 lexis in the expression of the advanced learner. *The Modern Language Journal*, *75*(4), 440–448. https://doi.org/10.2307/329493

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different?. *Applied Linguistics, 19*(2), 255–271. https://doi.org/10.1093/applin/19.2.255

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., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing, 16*(1), 33–51. https://doi.org/10.1177/026553229901600103

Laufer, B., & Paribakht, T. S. (1998). The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning*, *48*(3), 365–391. https://doi.org/10.1111/0023-8333.00046

Leech, D. (1994). Problematic ESL content word choice in writing: A proposed foundation of descriptive categories. *Issues in Applied Linguistics*, *5*(1), 83–102.

Llach, M. P. A., & Catalán, R. M. J. (2007). Lexical reiteration in EFL young learners' essays: does it relate to the type of instruction?. *International Journal of English Studies*, *7*(2), 85–104.

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

Macaro, E., Hultgren, A. K., Kirkpatrick, A., & Lasagabaster, D. (2019). English medium instruction: Global views and countries in focus: Introduction to the symposium held at the Department of Education, University of Oxford on Wednesday 4 November 2015. *Language Teaching*, *52*(2), 231–248. http://doi.org/10.1017/S0261444816000380

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Volume I: Transcription format and programs, volume II: The database*. Lawrence Erlbaum Associates.

Mahdi, H. S. (2018). Effectiveness of mobile devices on vocabulary learning: A meta-analysis. *Journal of Educational Computing Research*, *56*(1), 134–154. https://doi.org/10.1177/0735633117698826

Malvern, D. D., & Richards, B. J. (1997). A new measure of lexical diversity. *British Studies in Applied Linguistics*, *12*, 58–71.

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, *19*(1), 85–104. https://doi.org/10.1191/0265532202lt221oa

Malvern, D. D., Richards, B. J., Chipere, N., & Durán, P. (2004). *Lexical Diversity and Language Development: Quantification and Assessment*. Palgrave Macmillan.

Mass, H. D. (1972). Über den zusammenhang zwischen wortschatzumfang und länge eines textes. *Zeitschrift für Literaturwissenschaft und Linguistik*, *2*(8), 73–79.

Masrai, A. (2019). Vocabulary and reading comprehension revisited: Evidence for high-, mid-, and low-frequency vocabulary knowledge. *Sage Open*, *9*(2), 2158244019845182. https://doi.org/10.1177/2158244019845182

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 No. 3199485). [Doctoral dissertation, The University of Memphis]. ProQuest Dissertations. https://www.proquest.com/dissertations-theses/assessment-range-usefulness-lexical-diversity/docview/305349212/se-2

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

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

McCarthy, P. M., Watanabi S, & Lamkin, T. A. (2012). The Gramulator: A Tool to Identify Differential Linguistic Features of Correlative Text Types. In P. M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied Natural Language Processing: Identification, Investigation, and Resolution* (pp. 312–333), IGI Global.

McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*, *39*(6), 823–845. https://doi.org/10.1093/applin/amw050

McLean, S., & Kramer, B. (2015). The creation of a new vocabulary levels test. *Shiken*, *19*(2), 1–11.

McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*, *37*(3), 389–411. https://doi.org/10.1177/0265532219898380

Meara, P. (1984). The study of lexis in interlanguage. *Interlanguage*, 225–235.

Meara, P. (1990). A note on passive vocabulary. *Interlanguage Studies Bulletin (Utrecht)*, *6*(2), 150–154. https://doi.org/10.1177/026765839000600204

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.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*(2), 142–154. https://doi.org/10.1177/026553228700400202

Meara, P., & Fitzpatrick, T. (2000). Lex30: An improved method of assessing productive vocabulary in an L2. *System*, *28*(1), 19–30. https://doi.org/10.1016/S0346-251X(99)00058-5

Meara, P., & Jones, G. (1988). Vocabulary Size as a Placement Indicator. In P. Grunwell (Ed.), *Applied Linguistics in Society* (pp. 80–87). London.

Mendelsohn, D. J. (1981). We should assess lexical richness, not only lexical error. In *TESOL Convention, Detroit*.

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.), *L2 Vocabulary Acquisition,*

Knowledge and Use. New Perspectives on Assessment and Corpus Analysis (pp. 57–78). Eurosla.

Milton, J., & Alexiou, T. (2009). Vocabulary size and the common European framework of reference for languages. In B. Richards, 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). Springer.

Milton, J., Wade, J. & Hopkins, N. (2010). Aural word recognition and oral competence in English as a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse & M. Torreblanca-López (Eds.), *Insights into Non-native Vocabulary Teaching and Learning* (pp. 83–98). Multilingual Matters. https://doi.org/10.21832/9781847692900-007

Mochizuki, M., & Aizawa, K. (2000). An affix acquisition order for EFL learners: An exploratory study. *System*, *28*(2), 291–304. https://doi.org/10.1016/S0346-251X(00)00013-0

Myint Maw, T. M., Clenton, J., & Higginbotham, G. (2022). Investigating whether a flemma count is a more distinctive measurement of lexical diversity. *Assessing Writing*, *53*, 100640. https://doi.org/10.1016/j.asw.2022.100640

Nagy, W., & Scott, J. (2000). Vocabulary Processes. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson & R. Barr (Eds.). *Handbook of Reading Research* (pp. 269–284). Routledge.

Nagy, W., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly*, *47*(1), 91–108. https://doi.org/10.1002/RRQ.011

Nakata, T. (2011). Computer-assisted second language vocabulary learning in a paired-associate paradigm: A critical investigation of flashcard software. *Computer Assisted Language Learning*, *24*(1), 17–38. https://doi.org/10.1080/09588221.2010.520675

Nakata, T. (2020). Learning Words With Flash Cards and Word Cards. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 304–340). Routledge.

Nation, I. S. P. (1983) Testing and teaching vocabulary. *Guidelines*, *5*(1), 12–25.

Nation, I. S. P. (1984). *Vocabulary Lists*. Victoria University of Wellington, English Language Institute, Wellington, New Zealand.

Nation, I. S. P. (1990) *Teaching and Learning Vocabulary*. Newbury House.

Nation, I. S. P. (2001). *Learning vocabulary in another language (1st Edition)*. Cambridge University Press.

Nation, I. S. P. (2007). Fundamental issues in modelling and assessing vocabulary knowledge. In D. Helmut, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. 35–43). Cambridge University Press.

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

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. John Benjamins Publishing Company. https://doi.org/10.1075/z.208

Nation, I. S. P. (2017). The BNC/COCA level 6 word family lists (Version 2.0.0) [Data set]. http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

Nation, P. (2020). The different aspects of vocabulary knowledge. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 15–29). Routledge.

Nation, P. (2021). Thoughts on word families. *Studies in Second Language Acquisition*, *43*(5), 969–972. http://doi.org/10.1017/S027226312100067X

Nation, I. S. P. (2022). *Learning vocabulary in another language (3rd Edition).* Cambridge University Press.

Nation, P. (n.d.). *Vocabulary lists*. https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-lists

Nation, P., & Bauer, L. (2023). What is morphological awareness and how can you develop it? *Language Teaching Research*, *33*, 80–98. doi:10.32038/ltrq.2023.33.04

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.

Olinghouse, N. G., & Leaird, J. T. (2009). The relationship between measures of vocabulary and narrative writing quality in second-and fourth-grade students. *Reading and Writing*, *22*, 545–565. https://doi.org/10.1007/s11145-008-9124-z

Olinghouse, N. G., & Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. *Reading and Writing*, *26*, 45–65. https://doi.org/10.1007/s11145-012-9392-5

Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics*, *25*, 26–45. https://doi.org/10.1017/S0267190505000024

Ouelette, G. P. (2006). What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of Educational Psychology*, *98*(3), 554–566. https://doi.org/10.1037/0022-0663.98.3.554

Palmer, H. E. (1921). *The principles of language study*. Kessinger Publishing.

Paribakht, T. S., & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady & T. Huckin (Eds.), *Second Language Vocabulary Acquisition: A Rationale for Pedagogy* (pp. 174–200). Cambridge University Press.

Paul, P. V., Stallman, A. C. & O'Rourke, J. P. (1990). *Using three test formats to assess good and poor readers' word knowledge*. Technical Report No. 509, Center for the Study of Reading, University of Illinois at Urbana-Champaign, IL.

Pellicer-Sánchez, A. (2019). Examining second language vocabulary growth: Replications of Schmitt (1998) and Webb & Chang (2012). *Language Teaching*, *52*(4), 512–523. http://doi.org/10.1017/S026144481800037X

Pellicer-Sánchez, A. (2020). Learning single words vs. multiword items. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 158–173). Routledge.

Pinchbeck, G. G. (2014). Lexical frequencies profiling of Canadian high school diploma exam expository writing: L1 and L2 academic English. *Roundtable presentation at American Association of Applied Linguistics, Toronto, Ontario*.

Plonsky, L., Egbert, J., & Laflair, G. T. (2015). Bootstrapping in applied linguistics: Assessing its potential using shared data. *Applied Linguistics*, *36*(5), 591–610. https://doi.org/10.1093/applin/amu001

Qian, D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review*, *56*(2), 282–308. https://doi.org/10.3138/cmlr.56.2.282

Qian, D. D., & Lin, L. H. F. (2019). The relationship between vocabulary knowledge and language proficiency. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies*, (pp. 66–80). Routledge.

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

Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing*, *18*(1), 1–32. https://doi.org/10.1177/026553220101800101

Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly*, 77–89. https://doi.org/10.2307/3585941

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

Roquet, H., & Pérez-Vidal, C. (2017). Do productive skills improve in content and language integrated learning contexts? The case of writing. *Applied Linguistics*, *38*(4), 489–511. https://doi.org/10.1093/applin/amv050

Sasao, Y., & Webb, S. (2017). The word part levels test. *Language Teaching Research*, *21*(1), 12–30. https://doi.org/10.1177/1362168815586083

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing,* 44–49. Manchester: University of Manchester.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the SIGDAT Workshop at the Seventh Conference of the European*

Chapter of the Association for Computational Linguistics, 172–176. Association for Computational Linguistics.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge University Press.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329–363. https://doi.org/10.1177/1362168808089921

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. (2019). Understanding vocabulary acquisition, instruction, and assessment: A research agenda. *Language Teaching*, *52*(2), 261–274. https://doi.org/10.1017/S0261444819000053

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. http://doi:10.1017/S0272263197001022

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55–88. https://doi.org/10.1177/026553220101800103

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

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*(2), 139–152. https://doi.org/10.1080/09571730802389975

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, *31*(4), 577–607. https://doi.org/10.1017/S0272263109990039

Stoeckel, T., Ishii, T., & Bennett, P. (2020). Is the lemma more appropriate than the flemma as a word counting unit?. *Applied Linguistics*, *41*(4), 601–606. https://doi.org/10.1093/applin/amy059

Stoeckel, T., & McLean, S. (2022). The case for combining lexical and morphological text profiling: A response to Cobb. *Reading in a Foreign Language*, *34*(1), 172–183. http://hdl.handle.net/10125/67418

Templin, M. C. (1957). *Certain language skills in children*. University of Minnesota Press.

Teng, F. (2016). An in-depth investigation into the relationship between vocabulary knowledge and academic listening comprehension. *TESL-EJ*, *20*(2), 1–17. https://files.eric.ed.gov/fulltext/EJ1113907

Treffers-Daller, J. (2013). Measuring lexical diversity among L2 learners of French: an exploration of the validity of D, MTLD and HD-D as measures of language ability. In S. Jarvis & M. Daller (Eds.), *Vocabulary Knowledge. Human Ratings and Automated Measures* (pp. 79–103). John Benjamins Publishing Company. https://doi.org/10.1075/sibil.47.05ch3

Treffers-Daller, J., Parslow, P., & Williams, S. (2018). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, *39*(3), 302–327. https://doi.org/10.1093/applin/amw009

Treffers-Daller, J., Mukhopadhyay, L., Balasubramanian, A., Tamboli, V., & Tsimpli, I. (2022). How ready are Indian primary school children for English medium instruction? An analysis of the relationship between the reading skills of low-SES children, their oral vocabulary and English input in the classroom in government schools in India. *Applied Linguistics*, *43*(4), 746–775. https://doi.org/10.1093/applin/amac003

Trenkic, D., & Warmington, M. (2019). Language and literacy skills of home and international university students: How different are they, and does it matter?. *Bilingualism: Language and Cognition*, *22*(2), 349–365. https://doi.org/10.1017/S136672891700075X

Trochim, W. M. (2006). *The multitrait-multimethod matrix*. Research Methods Knowledge Base.

Uchihara, T., & Clenton, J. (2020). Investigating the role of vocabulary size in second language speaking ability. *Language Teaching Research*, *24*(4), 540–556. https://doi.org/10.1177/1362168818799371

Uchihara, T., & Saito, K. (2016). Exploring the relationship between productive vocabulary knowledge and second language oral ability. *The Language Learning Journal*, *47*(1), 64–75. https://doi.org/10.1080/09571736.2016.1191527

Vidal, K., & Jarvis, S. (2020). Effects of English-medium instruction on Spanish students' proficiency and lexical diversity in English. *Language Teaching Research*, *24*(5), 568–587. https://doi.org/10.1177/1362168818817945

Walters, J. (2012). Aspects of validity of a test of productive vocabulary: Lex30. *Language Assessment Quarterly*, *9*(2), 172–185. https://doi.org/10.1080/15434303.2011.625579

Ward, J., & Chuenjundaeng, J. (2009). Suffix knowledge: Acquisition and applications. *System*, *37*(3), 461–469. https://doi.org/10.1016/j.system.2009.01.004

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, *30*(1), 79–95. http://doi.org/10.1017/S0272263108080042

Webb, S. (2010). A corpus driven study of the potential for vocabulary learning through watching movies. *International Journal of Corpus Linguistics*, *15*(4), 497–519. https://doi.org/10.1075/ijcl.15.4.03web

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. A., & Chang, A. C. S. (2012). Second language vocabulary growth. *RELC Journal*, *43*(1), 113–126. https://doi.org/10.1177/0033688212439367

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL-International Journal of Applied Linguistics*, *168*(1), 33–69. https://doi.org/10.1075/itl.168.1.02web

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review*, *53*(1), 13–40. https://doi.org/10.3138/cmlr.53.1.13

West, M. (l953). *A General Service List of English Words.* Longman.

Wilkinson, R. (2013). English-medium instruction at a Dutch university: Challenges and pitfalls. In A. Doiz., D. Lasagabaster., & J. M. Sierra (Eds.), *English-medium Instruction at Universities: Global Challenges* (pp. 3–24). Multilingual Matters.

Yamamoto, Y. (2011). Bridging the gap between receptive and productive vocabulary size through extensive reading. *The Reading Matrix*, *11*(3), 226–242.

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, *31*(2), 236–259. https://doi.org/10.1093/applin/amp024

Yule, G. U. (1944). *The statistical study of literary vocabulary*. Cambridge University Press.

Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition*, *27*(4), 567–595. https://doi.org/10.1017/S0272263105050254

Zhang, L. J., & Anual, S. B. (2008). The role of vocabulary in reading comprehension: The case of secondary school students learning English in Singapore. *RELC Journal*, *39*(1), 51–76. https://doi.org/10.1177/0033688208091140

Zhang, P., & Graham, S. (2020). Learning vocabulary through listening: The role of vocabulary knowledge and listening proficiency. *Language Learning*, *70*(4), 1017–1053. https://doi.org/10.1111/lang.12411

Zhang, S., & Zhang, X. (2022). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, *26*(4), 696–725. https://doi.org/10.1177/1362168820913998

Zheng, Y., Zhang, Y., & Yan, Y. (2016). Investigating the practice of The Common European Framework of Reference for Languages (CEFR) outside Europe: A case study on the assessment of writing in English in China. British Council.

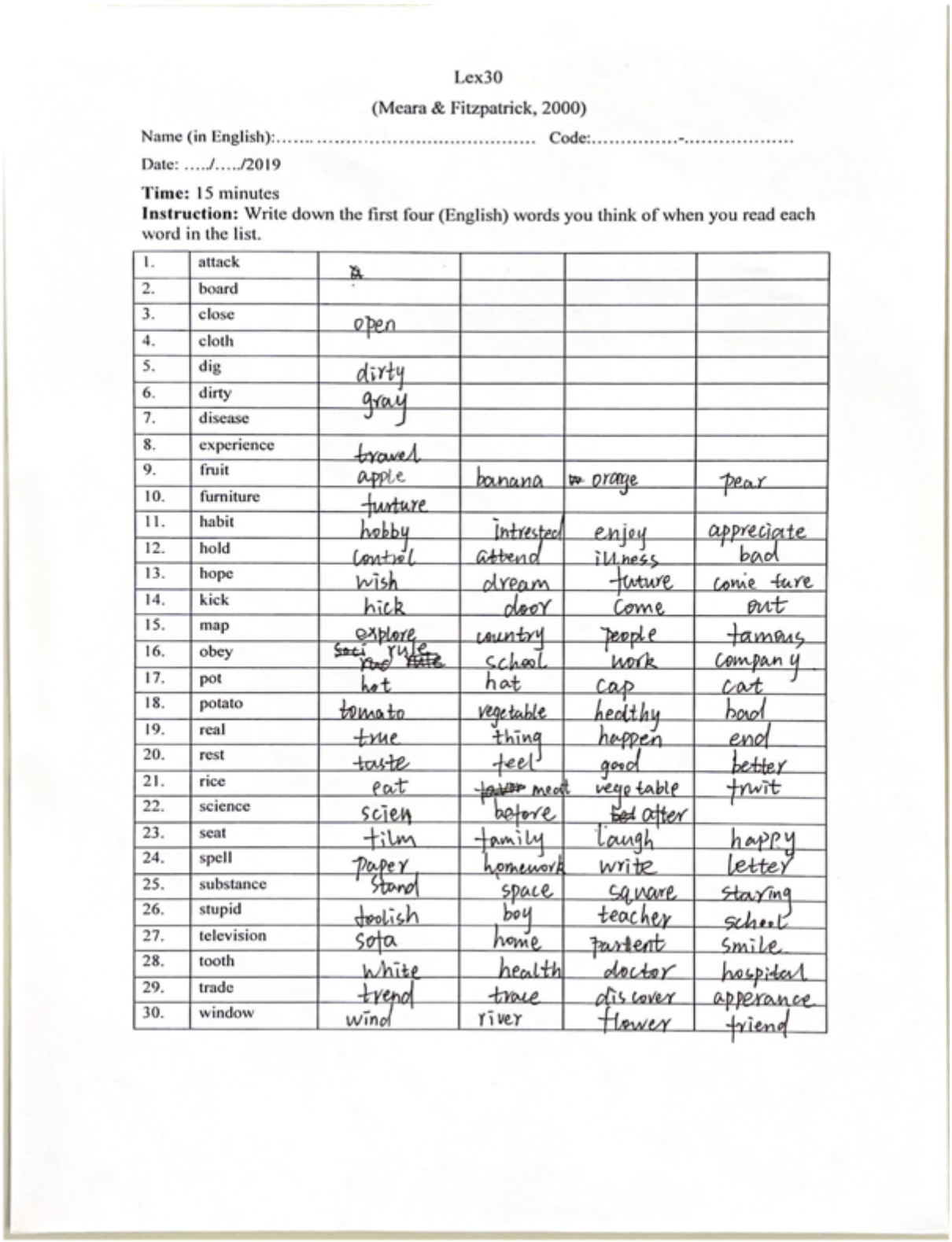
Zhong, H. & Hirsh D. (2009). Vocabulary growth in an English as a foreign language context. *University of Sydney Papers in TESOL*, *4*(4), 85–113.

Zipf, G. K. (1935). *The Psycho-Biology of Language*. Houghton Mifflin.

Appendix A:

Sample Responses of Vocabulary Tasks and IELTS Writing (Chapter Three)

Lex30 (Meara & Fitzpatrick, 2000)



Lex30

(Meara & Fitzpatrick, 2000)

Name (in English):..... Code:.....

Date: ...../...../2019

Time: 15 minutes

Instruction: Write down the first four (English) words you think of when you read each word in the list.

1.	attack				
2.	board				
3.	close	open			
4.	cloth				
5.	dig	dirty			
6.	dirty	gray			
7.	disease				
8.	experience	travel			
9.	fruit	apple	banana	orage	pear
10.	furniture	furture			
11.	habit	hobby	intrested	enjoy	appreciate
12.	hold	control	attend	illness	bad
13.	hope	wish	dream	future	come fure
14.	kick	hick	door	come	out
15.	map	explore	country	people	famous
16.	obey	~~Saci~~ ~~rod~~ rule ~~Hate~~	school	work	company
17.	pot	hot	hat	cap	cat
18.	potato	tomato	vegetable	healthy	bad
19.	real	true	thing	happen	end
20.	rest	taste	feel	good	better
21.	rice	eat	~~fa~~ meat	vege table	fruit
22.	science	scien	before	~~bef~~ after	
23.	seat	film	family	laugh	happy
24.	spell	paper	homework	write	letter
25.	substance	stand	space	square	staring
26.	stupid	foolish	boy	teacher	school
27.	television	sofa	home	parent	smile
28.	tooth	white	health	doctor	hospital
29.	trade	trend	trade	discover	apperance
30.	window	wind	river	flower	friend

G_Lex (Fitzpatrick & Clenton, 2017)

G-Lex
(Fitzpatrick & Clenton 2017)

Name (in English):..... Code:..... .....

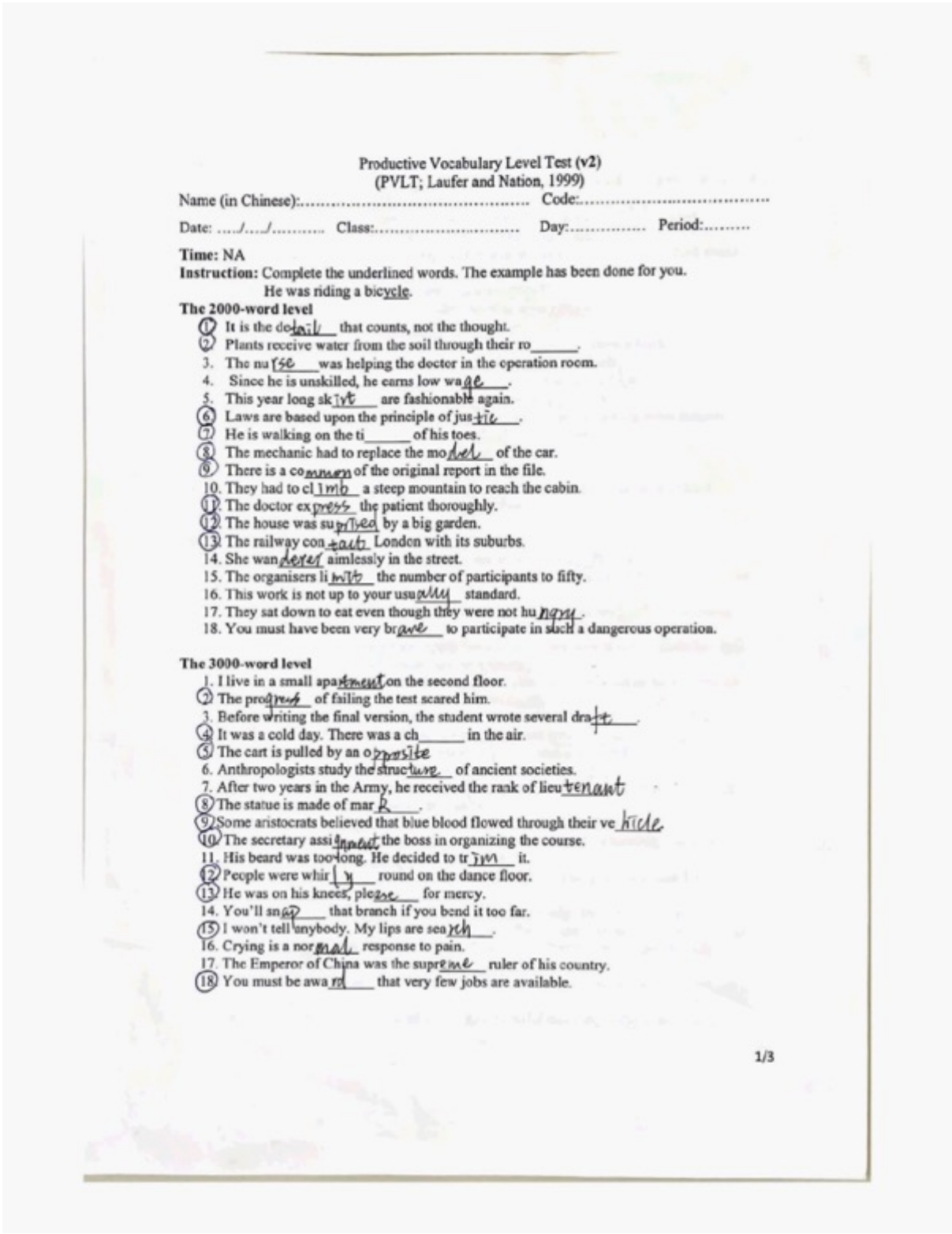
Date: ...../...../2019

Time: 15 minutes

Instruction: Write down five different words that might fit into each gap. The gaps are suitable for nouns, adjectives, and verbs in equal measure (eight sentences each).

1. She loved to _____ over the phone.	sing	speak	talk	read	express
2. When I feel sad I always go to the_____.	library	bed	room	playground	~~swim~~ gym
3. They think car-racing is_____.	wonderful	dangerous	excited	interesting	challenging
4. His colleague wanted to _____ the report.	publish	acquire	deliver	post	print
5. My favourite _____ is football.	ball	sport	thing	object	item
6. She looked _____ when she saw her friends.	beautiful	terrible	sad	awful	fearful
7. He couldn't _____ the car.	find	dive	have	use	possess
8. With a fire in my house I would save my ____.	family	mother	father	sister	brother
9. Many people feel ____about the environment.	sad	terrible	awful	fearful	comfortable
10. The parents _____ the children.	teach	love	hit	concern	kiss
11. He was happy with his _____.	face	gesture	expression	surface	~~look~~ appearance
12. He didn't think her teacher was _____ at all.	nothing	boring	brave	correct	severe
13. She always wanted to _____ after a busy day at work.	relax	sleep	study	read	talking
14. She sent _____to her mother.	flower	house	car	phone	necklace
15. The weather looked ____ before the game.	bad	sunny	cloudy	rainy	windy
16. He wanted to _____ the letter.	write	read	sing	throw	see
17. She was excited about _____.	film	story	book	poem	study
18. The girls thought the rock concert was____.	exhilarated	pleased	excited	successful	crazy
19. He took the chance to _____ the president.	hold	play	adequate	take	fun
20. He gave his boss _____.	~~love~~ suggestion	advice	idea	plan	choice
21. At the funeral the family felt ___.	grieved	heartbroken	sorrowful	sad	distressing
22. He always _____his breakfast.	eats	cooks	sloves	takes	buys
23. She put the food in the _____.	desk	table	chair	shelf	box
24. She was always ____ to those who needed help.	Concerned	bearing	sympathy	pity	care

Productive Vocabulary Level Test (the PVLT; Laufer & Nation, 1999)



Productive Vocabulary Level Test (v2)
(PVLT; Laufer and Nation, 1999)

Name (in Chinese):..... Code:.....

Date: ..../..../..... Class:..... Day:..... Period:.....

Time: NA

Instruction: Complete the underlined words. The example has been done for you.
He was riding a bicycle.

The 2000-word level

1. It is the de tail that counts, not the thought.
2. Plants receive water from the soil through their ro_____.
3. The nurse was helping the doctor in the operation room.
4. Since he is unskilled, he earns low wage.
5. This year long skirt are fashionable again.
6. Laws are based upon the principle of justice.
7. He is walking on the ti_____ of his toes.
8. The mechanic had to replace the model of the car.
9. There is a co mmon of the original report in the file.
10. They had to climb a steep mountain to reach the cabin.
11. The doctor express the patient thoroughly.
12. The house was surrived by a big garden.
13. The railway con nects London with its suburbs.
14. She wandered aimlessly in the street.
15. The organisers limit the number of participants to fifty.
16. This work is not up to your usually standard.
17. They sat down to eat even though they were not hungry.
18. You must have been very brave to participate in such a dangerous operation.

The 3000-word level

1. I live in a small apartment on the second floor.
2. The progress of failing the test scared him.
3. Before writing the final version, the student wrote several drafts.
4. It was a cold day. There was a ch_____ in the air.
5. The cart is pulled by an opposite
6. Anthropologists study the structure of ancient societies.
7. After two years in the Army, he received the rank of lieutenant
8. The statue is made of mar ble.
9. Some aristocrats believed that blue blood flowed through their ve ricle.
10. The secretary assi gnment the boss in organizing the course.
11. His beard was too long. He decided to trim it.
12. People were whir l y round on the dance floor.
13. He was on his knees, plea se for mercy.
14. You'll snap that branch if you bend it too far.
15. I won't tell anybody. My lips are sea rch.
16. Crying is a nor mal response to pain.
17. The Emperor of China was the supreme ruler of his country.
18. You must be awa re that very few jobs are available.

1/3

The 5000-word level

1. Some people find it difficult to become independent. Instead they prefer to be tied to their mother's ap propriate strings.
2. After finishing his degree, he entered upon a new ph enomenon in his career.
3. The workmen cleaned up the me ss before they left.
4. On Sunday, in his last se ntence in Church, the priest spoke against child abuse.
5. I saw them sitting on st ool at the bar drinking beer.
6. Her favorite musical instrument was a tru mpet.
7. The building is heated by a modern heating appa renie.
8. He received many com ments on his dancing skill.
9. People manage to buy houses by raising a mor tgage from a bank.
10. At the bottom of a blackboard there is a lo tter for chalk.
11. After falling off his bicycle, the boy was covered with bru she.
12. The child was holding a doll in her arms and hu g it.
13. We'll have to be inventive and de VIe a scheme for earning more money.
14. The picture looks nice; the colours bl end really well.
15. Nuts and vegetables are considered who lesome food.
16. The garden was full of fra grant flowers.
17. Many people feel depressed and gl oomy about the future of the mankind.
18. He is so depressed that he is cont emplate suicide.

The University Word List Level

1. I've had my eyes tested and the optician says my vi sion is good.
2. The anom alies of his position is that he is the chairman of the committee, but isn't allowed to vote.
3. In their geography class, the children are doing a special pro file on North America.
4. In a free country, people can apply for any job. They should not be discriminated against on the basis of colour, age, or s tatus .
5. A true dem onstrate should ensure equal rights and opportunities for all citizens.
6. The drug was introduced after medical res earch indisputably proved its effectiveness.
7. These courses should be taken in seq uence, not simultaneously.
8. Despite his physical condition, his int egrity was unaffected.
9. Governments often cut budgets in times of financial cri me .
10. The job offer sounded interesting at first. But when he realised what it would involve, his excitement subs tance gradually.
11. Research ind icate that men find it easier to give up smoking than women.
12. In a lecture, most of the talking is done by the lecturer. In a seminar, students are expected to part icularly in the discussion.
13. The airport is far away. If you want to ens ure that you catch your plane, you have to leave early.
14. It's difficult to ass ess a person's true knowledge by one or two tests.
15. The new manager's job was to res earch the company to its former profitability.
16. Even though the student didn't do well on the midterm exam, he got the highest mark on the fi nal le
17. His decision to leave home was not well thought out. It was not based on rat ional considerations.
18. The challenging job required a young, successful and dyn amic candidate.

The 10000-word level

1. The new vic tīmīzer was appointed by the bishop.
2. If your lips are sore, try lip sal iva, not medicine.
3. Much to his chag rin, he was not offered the job.
4. The actors exchanged ban_____ with reporters.
5. She wanted to marry nobility: a duke, a baron, or at least a vis a_____.
6. The floor in the ballroom was a mos aīc of pastel colours.
7. She has contributed a lot of money to various charities. She is known for her generosity and bene fiiinb
8. This is an unusual singer with a range of three oct ave.
9. A thro ttle controls the flow of gas into an engine.
10. Anyone found loo ming bombed houses and shops will be severly punished.
11. The crowd soon disp erse when the police arrived.
12. The wounded man squi rm on the floor in agony.
13. The dog crin klīng when it saw the snake.
14. He imme rse himself in a hot bubbly bath forgetting all his troubles for a moment.
15. The approaching storm stam ping the cattle into running wildly.
16. The problem is beginning to assume mam malia proportions.
17. His vind Icatin behaviour towards the thief was understandable.
18. He was arrested for illi cītly trading in drugs.

Vocabulary Levels Test (the VLT; Nation, 1983)

Name:_____ Age:_____ Date:___/___/2019

This is a vocabulary test. You must choose the right word to go with each meaning. Write the number of that word next to its meaning. Here is an example.

1	business	
2	clock	_____ part of a house
3	horse	___3___ animal with four legs
4	pencil	_____ something used for writing
5	shoe	
6	wall	

You answer it in the following way.

1	business	
2	clock	___6___ part of a house
3	horse	___3___ animal with four legs
4	pencil	___4___ something used for writing
5	shoe	
6	wall	

Some words are in the test to make it more difficult. You do not have to find a meaning for these words. In the example above, these words are business, clock, and shoe.

If you have no idea about the meaning of a word, do not guess. But if you think you might know the meaning, then you should try to find the answer.

Version 1 The 2,000 word level

1 birth			1 adopt		
2 dust	5	game	2 climb	2	go up
3 operation	6	winning	3 examine	5	look at closely
4 row	1	being born	4 pour	6	be on every side
5 sport			5 satisfy		
6 victory			6 surround		
1 choice			1 bake		
2 crop	6	heat	2 connect	2	join together
3 flesh	3	meat	3 inquire	6	walk without purpose
4 salary	4	money paid regularly for doing a job	4 limit	4	keep within a certain size
5 secret			5 recognize		
6 temperature			6 wander		
1 cap			1 burst		
2 education	2	teaching and learning	2 concern	1	break open
3 journey	5	numbers to measure with	3 deliver	5	make better
4 parent	2	going to a far place	4 fold	3	take something to someone
5 scale			5 improve		
6 trick			6 urge		
1 attack			1 original		
2 charm	6	gold and silver	2 private	1	first
3 lack	2	pleasing quality	3 royal	2	not public
4 pen	3	not having something	4 slow	6	all added together
5 shadow			5 sorry		
6 treasure			6 total		
1 cream			1 brave		
2 factory	1	part of milk	2 electric	6	commonly done
3 nail	6	a lot of money	3 firm	4	wanting food
4 pupil 5 3 1	4	person who is studying	4 hungry	1	having no fear
5 sacrifice			5 local		
6 wealth			6 usual		

Version 1 The 3,000 word level

1 belt
2 climate
3 executive
4 notion
5 palm
6 victim

4 idea
5 inner surface of your hand
1 strip of leather worn around the waist

1 acid
2 bishop
3 chill
4 ox
5 ridge
6 structure

3 cold feeling
4 farm animal
6 organization or framework

1 bench
2 charity
3 jar
4 mate
5 mirror
6 province

1 long seat
2 help to the poor
6 part of a country

1 boot
2 device
3 lieutenant
4 marble
5 phrase
6 vein

3 army officer
4 a kind of stone
6 tube through which blood flows

1 apartment
2 candle
3 draft
4 horror
5 prospect
6 timber

1 a place to live
5 chance of something happening
3 first rough form of something written

1 betray
2 dispose
3 embrace
4 injure
5 proclaim
6 scare

6 frighten
5 say publicly
4 hurt seriously

1 encounter
2 illustrate
3 inspire
4 plead
5 seal
6 shift

1 meet
4 beg for help
6 close completely

1 assist
2 bother
3 condemn
4 erect
5 trim
6 whirl

1 help
5 cut neatly
6 spin around quickly

1 annual
2 concealed
3 definite
4 mental
5 previous
6 savage

6 wild
3 clear and certain
1 happening once a year

1 dim
2 junior
3 magnificent
4 maternal
5 odd
6 weary

6 strange
3 wonderful
1 not clearly lit

Version 1 Academic Vocabulary

1 benefit
2 labor
3 percent
4 principle
5 source
6 survey

2 work
3 part of 100
6 general idea used to guide one's actions

1 element
2 fund 基金
3 layer
4 philosophy
5 proportion
6 technique

2 money for a special purpose
6 skilled way of doing something
4 study of the meaning of life

1 consent
2 enforcement
3 investigation
4 parameter
5 sum
6 trend

5 total
1 agreement or permission
3 trying to find information about something

1 decade
2 fee
3 file
4 incidence
5 perspective
6 topic

1 10 years
6 subject of a discussion
2 money paid for services

1 colleague
2 erosion
3 format
4 inclination
5 panel
6 violation

5 action against the law
4 wearing away gradually
3 shape or size of something

1 achieve
2 conceive
3 grant
4 link
5 modify
6 offset

5 change
4 connect together
1 finish successfully

1 convert
2 design
3 exclude
4 facilitate
5 indicate
6 survive

3 keep out
6 stay alive
1 change from one thing into another

1 anticipate
2 compile
3 convince
4 denote
5 manipulate
6 publish

5 control something skillfully
1 expect something will happen
2 produce books and newspapers

1 equivalent
2 financial
3 forthcoming
4 primary
5 random
6 visual

4 most important
6 concerning sight
2 concerning money

1 alternative
2 ambiguous
3 empirical
4 ethnic
5 mutual
6 ultimate

6 last or most important
1 something different that can be chosen
4 concerning people from a certain nation

Version 1 The 5,000 word level

1 balloon
2 federation
3 novelty
4 pail
5 veteran
6 ward

4 bucket 桶
3 unusual interesting thing
1 rubber bag that is filled with air

1 blend
2 devise
3 hug
4 lease
5 plague
6 reject

1 mix together
___ plan or invent
3 hold tightly in your arms

1 alcohol
2 apron
3 hip
4 lure
5 mess
6 phase

6 stage of development
5 state of untidiness or dirtiness
2 cloth worn in front to protect your clothes

1 abolish
2 drip
3 insert
4 predict
5 soothe
6 thrive

1 bring to an end by law
4 guess about the future
5 calm or comfort someone

1 apparatus
2 compliment
3 ledge
4 revenue
5 scrap
6 tile

2 expression of admiration
1 set of instruments or machinery
4 money received by the Government

1 bleed
2 collapse
3 precede 產3前.發生
4 reject
5 skip
6 tease

3 come before
2 fall down suddenly
5 move with quick steps and jumps

1 bulb
2 document
3 legion
4 mare
5 pulse
6 tub

4 female horse
3 large group of soldiers or people
2 a paper that provides information

1 casual
2 desolate
3 fragrant
4 radical
5 unique
6 wholesome

3 sweet-smelling
5 only one of its kind
6 good for your health

1 concrete
2 era
3 fiber
4 loop
5 plank
6 summit 頂点

4 circular shape
6 top of a mountain
2 a long period of time

1 gloomy 黑暗的
2 gross
3 infinite
4 limp
5 slim
6 vacant

6 empty
1 dark or sad
3 without end

Version 1 The 10,000 word level

1 antics	1	foolish behavior
2 batch	2	a group of things
3 connoisseur 行家	3	person with a good knowledge of art or music
4 foreboding		
5 haunch		
6 scaffold 支架		

1 auspices	4	confused mixture
2 dregs 渣滓	2	natural liquid present in the mouth
3 hostage 人质	3	worst and most useless parts of anything
4 jumble		
5 saliva		
6 truce		

1 casualty	1	someone killed or injured
2 flurry	6	being away from other people
3 froth	4	noisy and happy celebration
4 revelry 狂欢		
5 rut		
6 seclusion 隐居		

1 apparition 幽灵	1	ghost
2 botany	2	study of plants
3 expulsion	6	small pool of water
4 insolence		
5 leash		
6 puddle		

1 arsenal	4	happiness
2 barracks	5	difficult situation
3 deacon	6	minister in a church
4 felicity		
5 predicament		
6 spore		

1 acquiesce	2	to accept without protest
2 bask	1	sit or lie enjoying warmth
3 crease	3	make a fold on cloth or paper
4 demolish		
5 overhaul		
6 rape		

1 blaspheme	4	slip or slide
2 endorse	5	give care and food to
3 nurture	1	speak badly about God
4 skid		
5 squint		
6 straggle		

1 clinch	6	move very fast
2 jot	3	injure or damage
3 mutilate	4	burn slowly without flame
4 smolder		
5 topple		
6 whiz		

1 auxiliary	4	bad-tempered
2 candid	6	full of self-importance
3 luscious	1	helping, adding support
4 morose		
5 pallid		
6 pompous		

1 dubious	2	rude
2 impudent	6	very ancient
3 languid	4	of many different kinds
4 motley		
5 opaque		
6 primeval		

IELTS Writing Topic and Its Sample Response (Chapter Three)

ACADEMIC WRITING SAMPLE TASK 2B

You should spend about 40 minutes on this task.

Write about the following topic:

The threat of nuclear weapons maintains world peace. Nuclear power provides cheap and clean energy.

The benefits of nuclear technology far outweigh the disadvantages.

To what extent do you agree or disagree?

Give reasons for your answer and include any relevant examples from your knowledge or experience.

Write at least 250 words.

In today's increasingly lack of energy and rapid development of science and technology, the peaceful development and safe use of new energy is an important way to solve the energy crisis, such as the use of wind energy, nuclear weapons, solar energy.

Our use of nuclear weapons was originally a need for war. When peace was the mainstream of world. ~~He~~ a nuclear ~~weapons~~ energy for us was how to effectively use and develop.

According to reports from various countries, there were five main views on the explosion of China's first atomic bomb in international news at that time: 1. warning the leaders of the new soviet union to correct the course 2. Breaking the ~~he~~ nuclear monopoly of major powers and ending nuclear blackmail by major powers against backward countries is an important international force for nuclear balance. 3. To encourage the western countries to initiate dialogue with China, but also to return to the table to talk, talk well. 4. The internortional left is elated and the future looks bright 5. The imperialists and lackeys protested and began to change their diplomatic actions.

It should be said that there are all still many lessons we can draw from this history. such as how to handle relations between major countries, how to keep China in the most favorable international position, what position China was able to take and will take on the issue of nuclear weapons, and what should be paid attention to.

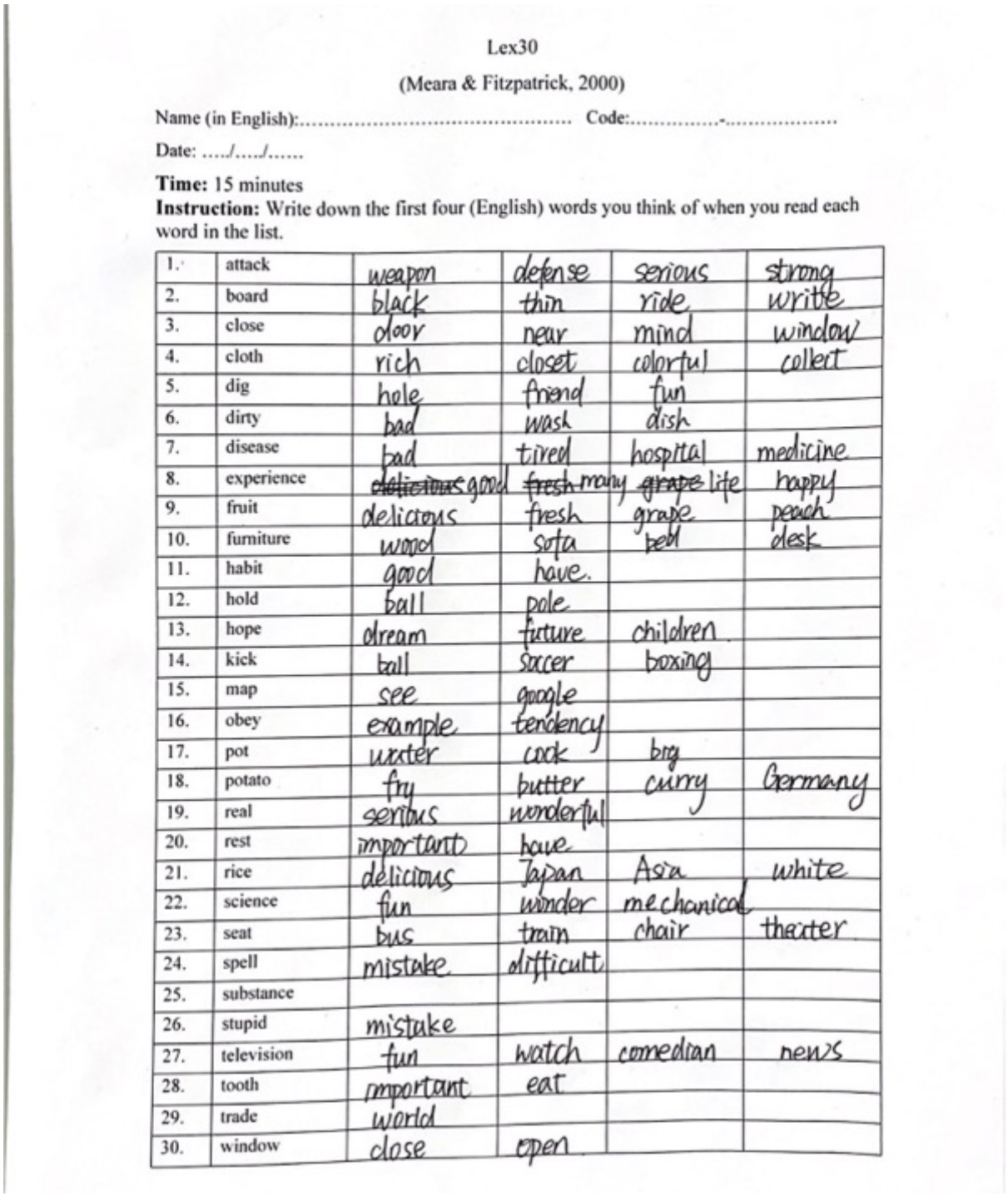
Traditional energy often refers to the fossil energy such as coal and oil and gas in low carbon ~~such~~ society under the advocate of nuclear power is the best choice for the first low cost in the second it is the third it is a kind of clean energy raw materials theoretically not face the status quo of energy depletion. of course, also have the danger of nuclear radition the need to strictly control the government. so overall benefits outweight the risk of can avoid (except for natural disasters. but to do the corresponding protection measures).

Appendix B:

Sample Responses of Vocabulary Tasks and IELTS Writing (Chapter Four)

Lex30 (Meara & Fitzpatrick, 2000)

Time: 15 minutes
Instruction: Write down the first four (English) words you think of when you read each word in the list.



Lex30

(Meara & Fitzpatrick, 2000)

Name (in English):..... Code:.....-.....

Date: ...../...../.....

Time: 15 minutes
Instruction: Write down the first four (English) words you think of when you read each word in the list.

1.	attack	weapon	defense	serious	strong
2.	board	black	thin	ride	write
3.	close	door	near	mind	window
4.	cloth	rich	closet	colorful	collect
5.	dig	hole	friend	fun	
6.	dirty	bad	wash	dish	
7.	disease	bad	tired	hospital	medicine
8.	experience	~~delicious~~ good	~~fresh~~ many	~~grape~~ life	happy
9.	fruit	delicious	fresh	grape	peach
10.	furniture	wood	sofa	bed	desk
11.	habit	good	have.		
12.	hold	ball	pole		
13.	hope	dream	future	children	
14.	kick	ball	soccer	boxing	
15.	map	see	google		
16.	obey	example	tendency		
17.	pot	water	cook	big	
18.	potato	fry	butter	curry	Germany
19.	real	serious	wonderful		
20.	rest	important	have.		
21.	rice	delicious	Japan	Asia	white
22.	science	fun	wonder	mechanical	
23.	seat	bus	train	chair	theater
24.	spell	mistake	difficult		
25.	substance				
26.	stupid	mistake			
27.	television	fun	watch	comedian	news
28.	tooth	important	eat		
29.	trade	world			
30.	window	close	open		

G_Lex (Fitzpatrick & Clenton, 2017)

Time: 15 minutes
Instruction: Write down five different words that might fit into each gap. The gaps are suitable for nouns, adjectives, and verbs in equal measure (eight sentences each).

G-Lex
(Fitzpatrick & Clenton 2017)

Name (in English):..... Code:.....-.....

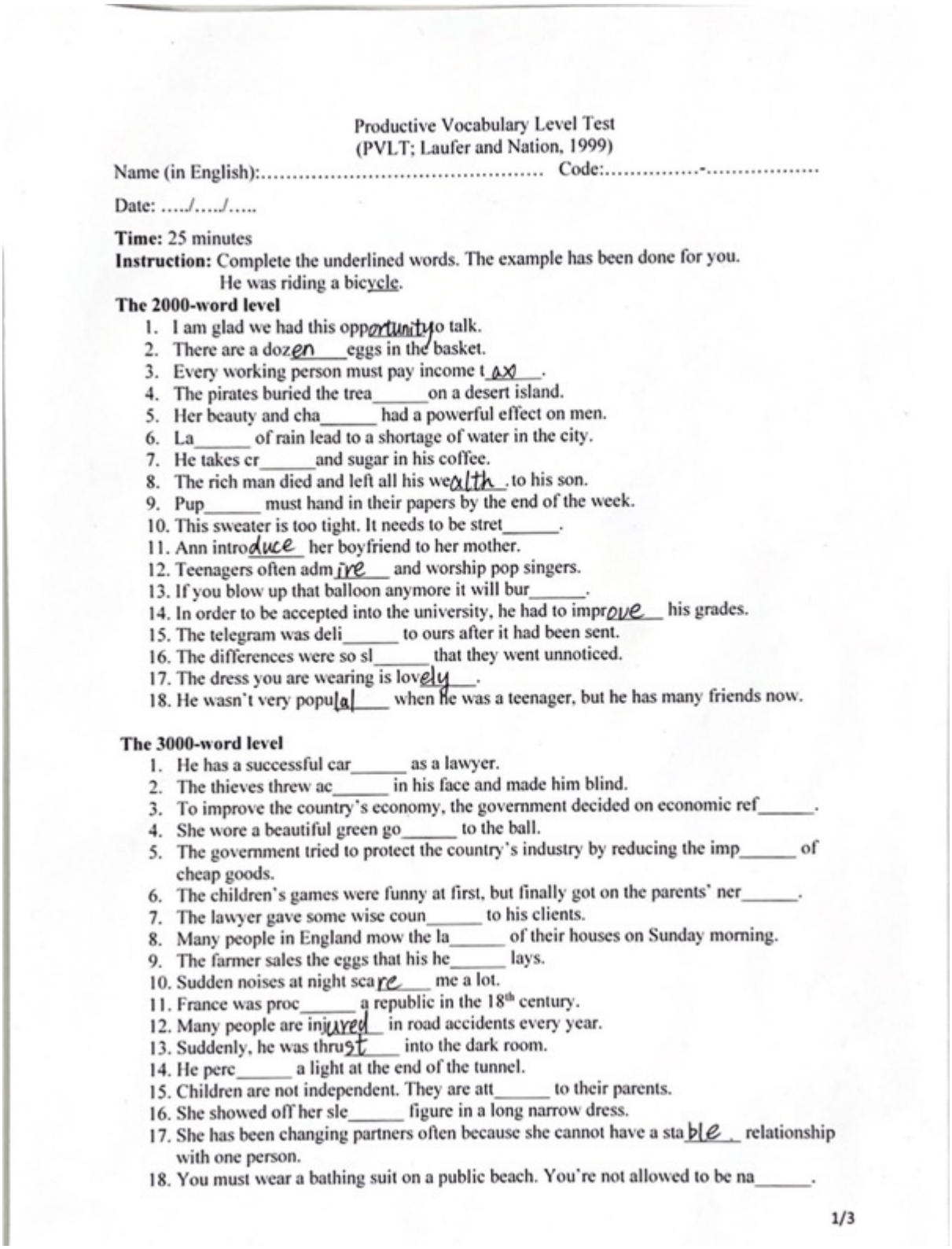
Date: ...../...../.....

Time: 15 minutes
Instruction: Write down five different words that might fit into each gap. The gaps are suitable for nouns, adjectives, and verbs in equal measure (eight sentences each).

1. She loved to _____ over the phone.	talk	look	get	buy	make
2. When I feel sad I always go to the_____.	university	hospital	school	city	place
3. They think car-racing is_____.	amazing	exciting	fast	wonderful	worldwide
4. His colleague wanted to _____ the report.	write	post	look	show	get
5. My favourite _____ is football.	sport	game	scene	ball	stadium
6. She looked _____ when she saw her friends.	happy	sad	enjoying	tired	upset
7. He couldn't _____ the car.	take	drive	buy	look	make
8. With a fire in my house I would save my ____.	life	dog	treasure	money	memory
9. Many people feel ____about the environment.	bad	dangerous	sad	angry	happy
10. The parents _____ the children.	talk to	look	bring up	show	take
11. He was happy with his _____.	success	examination	result	art	memory
12. He didn't think her teacher was _____ at all.	wise	good	bad	awful	stupid
13. She always wanted to _____ after a busy day at work.	travel	have a rest	read books		
14. She sent _____to her mother.	letter	gift			
15. The weather looked ____ before the game.	bad	good			
16. He wanted to _____ the letter.	receive	sent			
17. She was excited about _____.	sports	game			
18. The girls thought the rock concert was____.	cool	wonderful	excited		
19. He took the chance to _____ the president.	meet	talk to			
20. He gave his boss _____.	present				
21. At the funeral the family felt ___.					
22. He always _____his breakfast.	eat	look			
23. She put the food in the _____.	desk				
24. She was always ____ to those who needed help.					

Productive Vocabulary Level Test (the PVLT; Laufer & Nation, 1999)

Time: 25 minutes
Instruction: Complete the underlined words. The example has been done for you.
He was riding a bicycle.



Productive Vocabulary Level Test
(PVLT; Laufer and Nation, 1999)

Name (in English):..... Code:.....

Date: ...../...../.....

Time: 25 minutes
Instruction: Complete the underlined words. The example has been done for you.
He was riding a bicycle.

The 2000-word level

1. I am glad we had this opportunity to talk.
2. There are a dozen eggs in the basket.
3. Every working person must pay income tax.
4. The pirates buried the trea_____ on a desert island.
5. Her beauty and cha_____ had a powerful effect on men.
6. La_____ of rain lead to a shortage of water in the city.
7. He takes cr_____ and sugar in his coffee.
8. The rich man died and left all his wealth to his son.
9. Pup_____ must hand in their papers by the end of the week.
10. This sweater is too tight. It needs to be stret_____.
11. Ann introduce her boyfriend to her mother.
12. Teenagers often admire and worship pop singers.
13. If you blow up that balloon anymore it will bur_____.
14. In order to be accepted into the university, he had to improve his grades.
15. The telegram was deli_____ to ours after it had been sent.
16. The differences were so sl_____ that they went unnoticed.
17. The dress you are wearing is lovely.
18. He wasn't very popula when he was a teenager, but he has many friends now.

The 3000-word level

1. He has a successful car_____ as a lawyer.
2. The thieves threw ac_____ in his face and made him blind.
3. To improve the country's economy, the government decided on economic ref_____.
4. She wore a beautiful green go_____ to the ball.
5. The government tried to protect the country's industry by reducing the imp_____ of cheap goods.
6. The children's games were funny at first, but finally got on the parents' ner_____.
7. The lawyer gave some wise coun_____ to his clients.
8. Many people in England mow the la_____ of their houses on Sunday morning.
9. The farmer sales the eggs that his he_____ lays.
10. Sudden noises at night scare me a lot.
11. France was proc_____ a republic in the 18th century.
12. Many people are injured in road accidents every year.
13. Suddenly, he was thrust into the dark room.
14. He perc_____ a light at the end of the tunnel.
15. Children are not independent. They are att_____ to their parents.
16. She showed off her sle_____ figure in a long narrow dress.
17. She has been changing partners often because she cannot have a stable relationship with one person.
18. You must wear a bathing suit on a public beach. You're not allowed to be na_____.

1/3

The 5000-word level

1. Soldiers usually swear an oa _____ of loyalty to their country.
2. The voter placed the ball_____ in the box.
3. They keep their valuables in a vau_____ at the bank.
4. A bird perched at the window led_____.
5. The kitten is playing with a ball of ya_____.
6. The thieves have forced an entrance into the building.
7. The small hill was really a burial mou_____.
8. We decided to celebrate new year's e_____ together.
9. The solider was asked to choose between infinitary and cav_____.
10. This is a complex problem which is difficult to compr_____.
11. The angry crowd sho_____ the prisoner as he was leaving the court.
12. Don't pay attention to this rude remark. Just ign ore it.
13. The management held a secret meeting. The issues discussed were not disc_____ to the workers.
14. We could hear the sergeant bel_____ commands to the troops.
15. The boss got angry with the secretary and it took a lot of tact to soo_____ him.
16. We do not have adeq_____ information to make a decision.
17. She is not a child, but a mat_____ woman. She can make her own decisions.
18. The prisoner was put in soli_____ confinement.

The University Word List Level

1. There has been a recent tr_____ among prosperous families towards a smaller number of children.
2. The ar_____ of his office is 25 square meters.
3. Philosophy examines the meaning of life.
4. According to the communist doc_____, workers should rule the world.
5. Spending many years together deepened their inti_____.
6. He usually read the sport section of the newspaper first.
7. Because of the doctors' strike the clinic is closed today.
8. There are several misprints on each page of this te_____.
9. The suspect had both opportunity and mot_____ to commit the murder.
10. They insp_____ all products before sending them out to stores.
11. A considerable amount of evidence was accum_____ during the investigation.
12. The victim's shirt was satu_____with blood.
13. He is irresponsible. You can not re_____ on him for help.
14. It's impossible to evaluate these results without knowing about the research methods that were used.
15. He finally att_____ a position of power in the company.
16. The story tells us about a crime and subs_____ punishment.
17. In a hom_____ class all students are of a similar proficiency.
18. The urge to survive is inh_____ in all creatures.

The 10000-word level

1. The baby is wet. Her dia_____ needs changing.
2. The prisoner was released on par_____.
3. Second year University students in the US are called soph_____.
4. Her favorite flowers were or_____.
5. The insect causes damage to plants by its toxic sec_____.
6. The evac_____ of the building saved many lives.
7. For many people, wealth is a prospect of unimaginable felic_____.
8. She found herself in a pred_____ without any hope for a solution.
9. The deac_____ helped with the care of the poor of the parish.
10. The hurricane whi_____ along the coast.
11. Some coal was still smol_____ among the aches.
12. The dead bodies were muti_____ beyond recognition.
13. She was siting on a balcony and bas_____ in the sun.
14. For years waves of invaders pill_____ towns along the coast.
15. The rescue attempt could not proceed quickly. It was imp_____ by bad weather.
16. I wouldn't hire him. He is unmotivated and indo_____.
17. Computers have made typewriters old-fashioned and obs_____.
18. Watch out for his wil_____ tricks.

IELTS Writing Topic and Sample Response (Chapter Four)

You should spend about 40 minutes on this task.

Present a written case to an educated reader with no specialist knowledge of the following topic.

At the present time, the population of some countries includes a relatively large number of young adults, compared with the number of older people. Do the advantages of this situation outweigh the disadvantages?

You should use your own ideas, knowledge, and experience and support your argument with examples and relevant experience. Write at least 250 words.

In Japan today, the declining birthrate and aging population have become a social problem, and the government is lamenting the decline of the young adult, the labor force generation. In order to grow as a country in the future, it is necessary to have the power of the younger generation, so it is considered that the advantage is greater if there are more the number of young adults than them of older people.

There are several reasons why I think it's better to have a younger generation than the old people. First, Simply, the image of the country will be youthful and vibrant. Young people have more hope for the future than the elderly generation, and I think they have a great desire to improve, so if there are many such people in society, it is thought that the country will improve as well. I think that will naturally brighten people's minds and will be able to have hope for the future of the country. Second, the reason is that the stability of the labor force brings the stability of the social security. pension system and other security for old age. In modern Japan, where there are few younger generations, the future survival of this pension system is feared. The stable pension system is a sense of trust in the country and people can live with peace of mind

In contrast, there are disadvantages to the fact that there are many younger people. One of them is that the number of old people who have experience as a person is shallow, and the number of old people who should be used as reference is small, and the object to learn of people decreases.

However, when comparing the advantages and disadvantages, it is thought that the younger generation has a very good influence on the elderly generation.

Appendix C:

IELTS Task 2 Writing Band Descriptors (Public Version) (Chapter Five)

Band	Task Response	Coherence and Cohesion	Lexical Resource	Grammatical Range and Accuracy
0	Does not attend Does not attempt the task in any way Writes a totally memorised response			
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
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 organizational 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
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
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
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 overuse 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

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
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 overgeneralise 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
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 Skillfully 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
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 Skillfully 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’

Appendix D:

Ethical Convention

Dear Students,

I would like to ask for your help with a research project.

This research project will explore the vocabulary and IELTS writing that second-language learners (you) use in academic tasks. To do this we intend to use the vocabulary and written tasks that you submit in this course to make a corpus. This corpus will be analysed with measures of lexical diversity. The aim is to gain a greater understanding of what vocabulary students such as yourselves are able to produce. It is also hoped that this will lead to improvements in teaching materials.

Note that an early step in the analysis will be to anonymise the data; your name will therefore not appear in any reports or articles concerning this research. Another point to consider is that the analysis will be done after the pre-sessional course has finished, so it will not affect your grade in any way.

As the data will come from tasks that you will do anyway as part of your course, there is nothing extra for you to do (apart from completing this form).

- If you are happy for us to use your data in this project, please click below to say that you agree.
- If you do not wish to be part of this project, that is fine too; this will not affect your grade on the course (or future academic work) in any way.
- If later you decide to withdraw (you do not need to give a reason), send me an email and I will remove your data from the database.

Appendix E:

Sample Responses of Vocabulary Tasks and IELTS Writing (Chapter Five)

Japanese Participants' Sample Responses

Lex30 (Meara & Fitzpatrick, 2000)

Time: 15 minutes
Instruction: Write down the first four (English) words you think of when you read each word in the list.

Lex30

(Meara & Fitzpatrick, 2000)

Name (in English):..... Code:.....-.....

Date: ...../...../.....

Time: 15 minutes
Instruction: Write down the first four (English) words you think of when you read each word in the list.

1.	attack	train	car	people	
2.	board	gome boord	snow boord	white boord	
3.	close	stop	door	window	
4.	cloth	home	economics	clothes	cut
5.	dig				
6.	dirty				
7.	disease	bad			
8.	experience	development	life	important	grow
9.	fruit	opple	eat	orange	tree
10.	furniture	necessary	house	favorite	family
11.	habit				
12.	hold	hand	body		
13.	hope	want	wish	peace	
14.	kick	leg			
15.	map	university	world	city	way
16.	obey				
17.	pot	hot	tea	cup	keep
18.	potato	snack	fly		
19.	real	hard	rebbionet op	work	life
20.	rest	important			
21.	rice	white	small		
22.	science	difficult	signal		
23.	seat	car	train	plane	
24.	spell	English	Language		
25.	substance	science	something	air	earth
26.	stupid				
27.	television	program	interesting	news	
28.	tooth	clean	dentist		
29.	trade	country	fair	oil	ship
30.	window	open	close	breat	dangerous

G_Lex (Fitzpatrick & Clenton, 2017)

Time: 15 minutes
Instruction: Write down five different words that might fit into each gap. The gaps are suitable for nouns, adjectives, and verbs in equal measure (eight sentences each).

G-Lex
(Fitzpatrick & Clenton 2017)

Name (in English):..... Code:.....-.....

Date: ...../...../.....

Time: 15 minutes
Instruction: Write down five different words that might fit into each gap. The gaps are suitable for nouns, adjectives, and verbs in equal measure (eight sentences each).

1. She loved to _____ over the phone.	gamma	internet	SNS		
2. When I feel sad I always go to the_____.	favourite	shopping			
3. They think car-racing is_____.	excited	funny	dangerous		
4. His colleague wanted to _____ the report.	write	think			
5. My favourite _____ is football.	sport	club	circle		
6. She looked _____ when she saw her friends.	happy	sad	angry		
7. He couldn't _____ the car.	drive	buy	rent	stop	
8. With a fire in my house I would save my ____.	baby	family	money	photograph	
9. Many people feel ____about the environment.	great	beautiful	dangerous	afraid	
10. The parents _____ the children.	help	grow	hotel	keep	
11. He was happy with his _____.	family	brother	friend	sister	girlfriend
12. He didn't think her teacher was _____ at all.	busy	sad	angry		
13. She always wanted to _____ after a busy day at work.	sleep	eat	go shopping		
14. She sent _____to her mother.	gift	food	salary		
15. The weather looked ____ before the game.	fine	bad	sunny	cloudy	raining
16. He wanted to _____ the letter.	write	read	sent		
17. She was excited about _____.	games	books	sports	songs	TV
18. The girls thought the rock concert was____.	excited	good	afraid		
19. He took the chance to _____ the president.	get	kill	meet	talk	replace
20. He gave his boss _____.	idea	hand out	gift		
21. At the funeral the family felt ___.	good	bad			
22. He always _____his breakfast.	have	eat	buy	give	throw
23. She put the food in the _____.	table	desk	floor	chair	
24. She was always ____ to those who needed help.	sad	poor	angry		

Productive Vocabulary Level Test (the PVLT; Laufer & Nation, 1999)

Time: 25 minutes
Instruction: Complete the underlined words. The example has been done for you.
He was riding a bicycle.

Productive Vocabulary Level Test
(PVLT; Laufer and Nation, 1999)

Name (in English):..... Code:.....-.....

Date: ...../...../.....

Time: 25 minutes
Instruction: Complete the underlined words. The example has been done for you.
He was riding a bicycle.

The 2000-word level

1. I am glad we had this opportunity to talk.
2. There are a doz_____ eggs in the basket.
3. Every working person must pay income t_____.
4. The pirates buried the trea_____ on a desert island.
5. Her beauty and cha_____ had a powerful effect on men.
6. La_____ of rain lead to a shortage of water in the city.
7. He takes cream and sugar in his coffee.
8. The rich man died and left all his we_____ to his son.
9. Pup_____ must hand in their papers by the end of the week.
10. This sweater is too tight. It needs to be stret_____.
11. Ann introduce her boyfriend to her mother.
12. Teenagers often adm_____ and worship pop singers.
13. If you blow up that balloon anymore it will bur_____.
14. In order to be accepted into the university, he had to improve his grades.
15. The telegram was deli_____ to ours after it had been sent.
16. The differences were so sl_____ that they went unnoticed.
17. The dress you are wearing is lovely.
18. He wasn't very popu lar when he was a teenager, but he has many friends now.

The 3000-word level

1. He has a successful car_____ as a lawyer.
2. The thieves threw acid in his face and made him blind.
3. To improve the country's economy, the government decided on economic ref_____.
4. She wore a beautiful green go_____ to the ball.
5. The government tried to protect the country's industry by reducing the imp_____ of cheap goods.
6. The children's games were funny at first, but finally got on the parents' ner_____.
7. The lawyer gave some wise coun_____ to his clients.
8. Many people in England mow the la_____ of their houses on Sunday morning.
9. The farmer sales the eggs that his he_____ lays.
10. Sudden noises at night scared me a lot.
11. France was proc_____ a republic in the 18th century.
12. Many people are inj_____ in road accidents every year.
13. Suddenly, he was thru_____ into the dark room.
14. He perc_____ a light at the end of the tunnel.
15. Children are not independent. They are attached to their parents.
16. She showed off her sle_____ figure in a long narrow dress.
17. She has been changing partners often because she cannot have a sta_____ relationship with one person.
18. You must wear a bathing suit on a public beach. You're not allowed to be na_____.

1/3

The 5000-word level

1. Soldiers usually swear an oa _____ of loyalty to their country.
2. The voter placed the ball_____ in the box.
3. They keep their valuables in a vau_____ at the bank.
4. A bird perched at the window led_____.
5. The kitten is playing with a ball of ya_____.
6. The thieves have forced an entrance into the building.
7. The small hill was really a burial mou_____.
8. We decided to celebrate new year's e_____ together.
9. The solider was asked to choose between infinitary and cav_____.
10. This is a complex problem which is difficult to compr_____.
11. The angry crowd sho_____ the prisoner as he was leaving the court.
12. Don't pay attention to this rude remark. Just ign_____ it.
13. The management held a secret meeting. The issues discussed were not disc_____ to the workers.
14. We could hear the sergeant bel_____ commands to the troops.
15. The boss got angry with the secretary and it took a lot of tact to soo_____ him.
16. We do not have adeq_____ information to make a decision.
17. She is not a child, but a mat_____ woman. She can make her own decisions.
18. The prisoner was put in soli_____ confinement.

The University Word List Level

1. There has been a recent tr_____ among prosperous families towards a smaller number of children.
2. The area of his office is 25 square meters.
3. Philosophy examines the meaning of life.
4. According to the communist doc_____, workers should rule the world.
5. Spending many years together deepened their inti_____.
6. He usually read the sport sec_____ of the newspaper first.
7. Because of the doctors' strike the cli_____ is closed today.
8. There are several misprints on each page of this te_____.
9. The suspect had both opportunity and mot_____ to commit the murder.
10. They insp_____ all products before sending them out to stores.
11. A considerable amount of evidence was accum_____ during the investigation.
12. The victim's shirt was satu_____with blood.
13. He is irresponsible. You can not re_____ on him for help.
14. It's impossible to eva_____ these results without knowing about the research methods that were used.
15. He finally att_____ a position of power in the company.
16. The story tells us about a crime and subs_____ punishment.
17. In a hom_____ class all students are of a similar proficiency.
18. The urge to survive is inh_____ in all creatures.

The 10000-word level

1. The baby is wet. Her dia_____ needs changing.
2. The prisoner was released on par_____.
3. Second year University students in the US are called soph_____.
4. Her favorite flowers were or_____.
5. The insect causes damage to plants by its toxic sec_____.
6. The evac_____ of the building saved many lives.
7. For many people, wealth is a prospect of unimaginable felic_____.
8. She found herself in a pred_____ without any hope for a solution.
9. The deac_____ helped with the care of the poor of the parish.
10. The hurricane whi_____ along the coast.
11. Some coal was still smol_____ among the aches.
12. The dead bodies were muti_____ beyond recognition.
13. She was siting on a balcony and bas_____ in the sun.
14. For years waves of invaders pill_____ towns along the coast.
15. The rescue attempt could not proceed quickly. It was imp_____ by bad weather.
16. I wouldn't hire him. He is unmotivated and indo_____.
17. Computers have made typewriters old-fashioned and obs_____.
18. Watch out for his wil_____ tricks.

IELTS Writing Topic and Sample Response (Chapter Five)

IELTS topic one:

You should spend about 40 minutes on this task.

Present a written case to an educated reader with no specialist knowledge of the following topic.

At the present time, the population of some countries includes a relatively large number of young adults, compared with the number of older people. Do the advantages of this situation outweigh the disadvantages?

You should use your own ideas, knowledge, and experience and support your argument with examples and relevant experience. Write at least 250 words.

In recent years, young people have a large percentage of the population than older people in many countries in the world, in contrast Japan is aging and the percentage of elderly is increasing. Various problems are occurring due to the high percentage of elderly people in Japan. What are the advantages and disadvantages of a country that has a high proportion of young people compared to Japan with many elderly people?

What I came up with about the advantage of a country with many young people was the working population and the large number of children. A large number of prime young people means a large working population, which may Lead to further economic and social development. Also, many young women will be able to born to children, which will increase the birth rate and increase number of children. Unlike these countries, Japan is suffering from a small working population and a declining birth rate. These countries are much better than Japan like this in terms of present and future workforce.

The disadvantages of these countries are that they have a large working population and a large number of children, making it difficult for them to find employment. If the number of employment is smaller than the working population, it will be difficult to find employment, and some young people will not be able to find employment. In recent years, women have advanced into society, so if there are many children, they will have to work and raise children, increasing the burden of parents. This may Lead to child abandonment and divorce. If these increase in the country, it will be a big problem. These countries and Japan are common in that raising children has became more difficult with the advancement of women in society. Even in Japan with declining birthrates, the difficulty of raising children and the anxiety of divorce of dual working parents are problems, so it is thought that it will be a country with more children in Japan.

I think the advantages are greater than the disadvantages when comparing the advantages and disadvantages of countries with a high population of young people. I think the difficulty of finding employment and the difficulty of raising children are also important issues, but I think the working population and the large number of children will Lead to social and economic stability both now and in the future. Even in Japan, where ~~are also important issues,~~ the population is aging and declining, the difficulty of childcare and the number of divorces are problems. That is why, unlike Japan, there are many children who support the working population and future society, so as to balance and economy are guaranteed. I think the advantage of having a large workforce and a large number of children with a high percentage of young people are important to the country because I live in Japan where employment, marriage, child, and retirement are problematic.

IELTS topic two:

You should spend about 40 minutes on this task.

Present a written argument or case to an educated reader with no specialist knowledge of the following topic. *It is necessary for parents to attend a parenting training course to bring their children up. Do you agree or disagree?*

You should use your own ideas, knowledge, and experience and support your arguments with examples and relevant experience. Write at least 250 words.

In recent years, some parents in Japan attend a parenting training course to bring their children up. Some may find it necessary for their parents to attend in this course, while others may not. I think it is necessary for parents to attend such events that support child-rearing, so I want to give need along with the reason.

In the first place, such as an education course for child-rearing has been established in recent years, and it is thought to the result of the advancement of nuclear families. How to raise children was natural teach parents, but it has become a rare practice in recent years when living away from parents has become commonplace. I think this is the only course that parents can learn properly about raising children in the age of nuclear families. I think it is necessary for children to learn how to raise children properly, as there are many problem of abandonment of parenting.

In addition, since many parents in the midst of child care gather in such a parenting training course, parents who are raising children can get to know each other. I think there are many parents who are worried about raising children and want someone to sympathize with the difficulty of raising children. In such a case, I think that it would be a great support for many parents to be able to meet people who have the same worries and sympathy with the difficult in this course. Being able to get to know and support each other is also major benefit of this course, and I think this is the most necessary encounter for parents during child care.

Some people believe that a parenting training course is no longer necessary if they are taught by parents following past practices; however, I think that the need for a parenting training course is increasing because the number of homes that do not live with parents has increased due to the nuclear family, and parents are not able to easily teach child-raising. I think it is necessary for modern society to increase the number of events in which parents can correctly learn about parenting and receive support, as in a parenting training course.

French Participants' Sample Responses (Chapter Five)

Lex30 (Meara & Fitzpatrick, 2000)

Time: 15 minutes
Instruction: Write down the first four (English) words you think of when you read each word in the list.

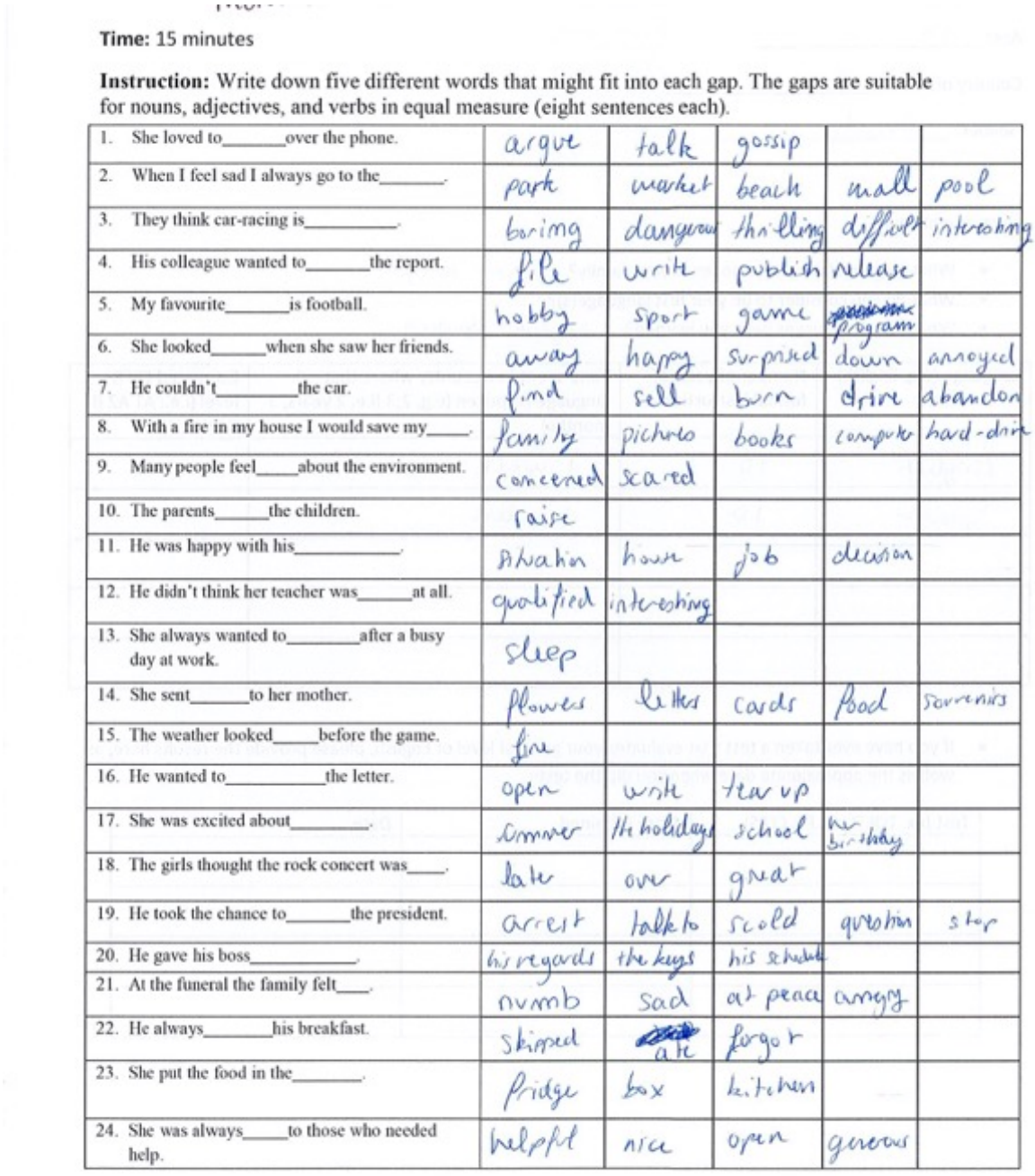
Time: 15 minutes

Instruction: Write down the first four (English) words you think of when you read each word in the list.

1.	attack	violent	invasion	physical	harmful
2.	board	cutting - board	surfboard	blackboard	
3.	close	near			
4.	cloth	fabric			
5.	dig	grave	shovel	deep	
6.	dirty	clean	stain		
7.	disease	illness	cure	medecine	symptoms
8.	experience	perspective	difficult	personal	
9.	fruit	apple	raw	sweet	
10.	furniture	bed	table	decor	
11.	habit	usual	repetitive	addiction	
12.	hold	tight			
13.	hope	future	optimistic	perspective	
14.	kick	harm	ball		
15.	map	roads	directions		
16.	obey	rules	power		
17.	pot	cook	pan	stove	
18.	potato	mash	fries		
19.	real	authentic	believable	factual	Reverse
20.	rest	quiet	sleep sleep	calm	
21.	rice	beans	food		
22.	science	facts	curiosity		
23.	seat	car			
24.	spell	curse			
25.	substance	container			
26.	stupid	dump	idiot	ignorant	dangerous
27.	television	entertainment	information	pop - culture	
28.	tooth	mouth	jaw	gums	tongue
29.	trade	exchange	financial		
30.	window	door	curtain	blinds	light

G_Lex (Fitzpatrick & Clenton, 2017)

Time: 15 minutes
Instruction: Write down five different words that might fit into each gap. The gaps are suitable for nouns, adjectives, and verbs in equal measure (eight sentences each).



Time: 15 minutes

Instruction: Write down five different words that might fit into each gap. The gaps are suitable for nouns, adjectives, and verbs in equal measure (eight sentences each).

1. She loved to_____over the phone.	argue	talk	gossip		
2. When I feel sad I always go to the_____.	park	market	beach	mall	pool
3. They think car-racing is_____.	boring	dangerous	thrilling	difficult	interesting
4. His colleague wanted to_____the report.	file	write	publish	release	
5. My favourite_____is football.	hobby	sport	game	~~program~~	
6. She looked_____when she saw her friends.	away	happy	surprised	down	annoyed
7. He couldn't_____the car.	find	sell	burn	drive	abandon
8. With a fire in my house I would save my_____.	family	pictures	books	computer	hard-drive
9. Many people feel_____about the environment.	concerned	scared			
10. The parents_____the children.	raise				
11. He was happy with his_____.	Situation	house	job	decision	
12. He didn't think her teacher was_____at all.	qualified	interesting			
13. She always wanted to_____after a busy day at work.	sleep				
14. She sent_____to her mother.	flowers	letters	cards	food	souvenirs
15. The weather looked_____before the game.	fine				
16. He wanted to_____the letter.	open	write	tear up		
17. She was excited about_____.	summer	the holidays	school	her birthday	
18. The girls thought the rock concert was_____.	late	over	great		
19. He took the chance to_____the president.	arrest	talk to	scold	question	star
20. He gave his boss_____.	his regards	the keys	his schedule		
21. At the funeral the family felt_____.	numb	sad	at peace	angry	
22. He always_____his breakfast.	skipped	~~ate~~	forgot		
23. She put the food in the_____.	fridge	box	kitchen		
24. She was always_____to those who needed help.	helpful	nice	open	generous	

Productive Vocabulary Level Test (the PVLT; Laufer & Nation, 1999)

Time: 25 minutes
Instruction: Complete the underlined words. The example has been done for you.
He was riding a bicycle.

CODE: MONT-T1-22
Time: 25-30 minutes

Instruction: Complete the underlined words. The example has been done for you.

EXAMPLE: He was riding a bic_____. → He was riding a bicycle.

The 2000-word level

1. I am glad we had this opp**ortunity** to talk.
2. There are a doz**en** eggs in the basket.
3. Every working person must pay income t**axes**.
4. The pirates buried the trea**sure** on a desert island.
5. Her beauty and cha**rm** had a powerful effect on men.
6. La**ck** of rain lead to a shortage of water in the city.
7. He takes cr**eam** and sugar in his coffee.
8. The rich man died and left all his we**alth** to his son.
9. Pup**ils** must hand in their papers by the end of the week.
10. This sweater is too tight. It needs to be stret**ched out**.
11. Ann intro**duced** her boyfriend to her mother.
12. Teenagers often adm**ire** and worship pop singers.
13. If you blow up that balloon anymore it will bur**st**.
14. In order to be accepted into the university, he had to impr**ove** his grades.
15. The telegram was deli**vered** to ours after it had been sent.
16. The differences were so sl**ight** that they went unnoticed.
17. The dress you are wearing is lov**ely**.
18. He wasn't very popu**lar** when he was a teenager, but he has many friends now.

The 3000-word level

1. He has a successful car**rier** as a lawyer.
2. The thieves threw ac**id** in his face and made him blind.
3. To improve the country's economy, the government decided on economic ref**orm**.
4. She wore a beautiful green go**wn** to the ball.
5. The government tried to protect the country's industry by reducing the imp**ortation** of cheap goods.
6. The children's games were funny at first, but finally got on the parents' ner**ves**.
7. The lawyer gave some wise coun**seling** to his clients.
8. Many people in England mow the la**wn** of their houses on Sunday morning.
9. The farmer sells the eggs that his he**ns** lays.
10. Sudden noises at night sca**re** me a lot.
11. France was proc**laimed** a republic in the 18th century.
12. Many people are inj**ured** in road accidents every year.

CODE :

13. Suddenly, he was thrusted into the dark room.
14. He perceived a light at the end of the tunnel.
15. Children are not independent. They are attached to their parents.
16. She showed off her slender figure in a long narrow dress.
17. She has been changing partners often because she cannot have a stable relationship with one person.
18. You must wear a bathing suit on a public beach. You're not allowed to be naked.

The 5000-word level

1. Soldiers usually swear an oath of loyalty to their country.
2. The voter placed the ballot in the box.
3. They keep their valuables in a vau at the bank.
4. A bird perched at the window led.
5. The kitten is playing with a ball of yarn.
6. The thieves have forced an entry into the building.
7. The small hill was really a burial mountain.
8. We decided to celebrate new year's eve together.
9. The solider was asked to choose between infantry and cav.
10. This is a complex problem which is difficult to comprehend.
11. The angry crowd shouted at the prisoner as he was leaving the court.
12. Don't pay attention to this rude remark. Just ignore it.
13. The management held a secret meeting. The issues discussed were not diss to the workers.
14. We could hear the sergeant bellow out commands to the troops.
15. The boss got angry with the secretary and it took a lot of tact to soothe him.
16. We do not have adequate information to make a decision.
17. She is not a child, but a mature woman. She can make her own decisions.
18. The prisoner was put in soli tary confinement.

The University Word List Level

1. There has been a recent tr among prosperous families towards a smaller number of children.
2. The area of his office is 25 square meters.
3. Philosopher examines the meaning of life.
4. According to the communist documents, workers should rule the world.
5. Spending many years together deepened their intimacy.
6. He usually read the sport section of the newspaper first.
7. Because of the doctors' strike the clinic is closed today.
8. There are several misprints on each page of this te.
9. The suspect had both opportunity and motive to commit the murder.

CODE :

10. They insp_____all products before sending them out to stores.
11. A considerable amount of evidence was accum_____during the investigation.
12. The victim's shirt was satu_____with blood.
13. He is irresponsible. You cannot re_____on him for help.
14. It's impossible to eva_____these results without knowing about the research methods that were used.
15. He finally att_____a position of power in the company.
16. The story tells us about a crime and subs_____punishment.
17. In a hom_____class all students are of a similar proficiency.
18. The urge to survive is inh_____in all creatures.

The 10000-word level

1. The baby is wet. Her dia_____needs changing.
2. The prisoner was released on par_____.
3. Second year University students in the US are called soph_____.
4. Her favorite flowers were or_____.
5. The insect causes damage to plants by its toxic sec_____.
6. The evac_____of the building saved many lives.
7. For many people, wealth is a prospect of unimaginable felic_____.
8. She found herself in a pred_____without any hope for a solution.
9. The deac_____helped with the care of the poor of the parish.
10. The hurricane whi_____along the coast.
11. Some coal was still smol_____among the ashes.
12. The dead bodies were muti_____beyond recognition.
13. She was sitting on a balcony and bas_____in the sun.
14. For years waves of invaders pill_____towns along the coast.
15. The rescue attempt could not proceed quickly. It was imp_____by bad weather.
16. I wouldn't hire him. He is unmotivated and indo_____.
17. Computers have made typewriters old-fashioned and obs_____.
18. Watch out for his wil_____tricks.

IELTS Writing Topic and Sample Response (Chapter Five)

IELTS topic one:

Academic writing sample task 2B

You should spend about 40 minutes on this task.

Write about the following topic:

The threat of nuclear weapons maintains world peace. Nuclear power provides cheap and clean energy. The benefits of nuclear technology far outweigh the disadvantages. To what extent do you agree or disagree?

Give reasons for your answer and include any relevant examples from your knowledge or experience. Write at least 250 words.

MONT-TI-22

The idea that nuclear energy overall benefits us more than it does harm is controversial. There is, of course some truth to the statement but nuclear weapons and technology are still a sensitive subject.

When it comes to nuclear weapons, advocates argue that its threat is what maintains world peace, in other words, we have seen the harm it can cause and don't want to repeat those events. Nuclear weapons have undeniably acted as the main tool to expose us to our own cruelty as a species, how far we are able to go in a conflict. shows us

Still, we can't erase the damage that has already been done and shouldn't try to justify these events by making them completely an example. Again, although our use of nuclear weapons has indeed served as a lesson, it was merely a tool. It is our own cruelty and lack of boundaries we should be afraid of, because if we ever come back to that point, I am sure we will find other resource to do as much, if not more damage.

The question of nuclear energy is even more complicated because of how little information we have. Sure, it does seem like a great solution to today's environmental issues, but we cannot be certain of the impact at the moment negative

it could have in the future; which does match the pattern ... we've been following: finding a quick solution to an issue and having to deal with the consequences later on. Thus, we are torn between two questions; should we think further ahead but do we have the time and the option to do so? We need to evaluate our options with the limited time we have to actually make changes, which some would argue, is none. I believe the man ... would be committing to a single source of energy, thus having no possible back-up options if the source, mistake, proves to be unreliable. Also, since we do already know the issue nuclear waste represents, it would obviously be unwise to produce enormous amounts. Still, nuclear technology is an interesting option, to be considered with caution.

IELTS topic two:

Writing about the following topic:

The first car appeared on British roads in 1888. By the year 2020 there may be as many as 35 million vehicles on British roads. Alternative forms of transport should be encouraged and international laws introduced to control car ownership and use. To what extent do you agree or disagree?

Give reasons for your answer and include any relevant examples from your knowledge or experience. Write at least 250 words.

Write about the following topic:

> ***The first car appeared on British roads in 1888. By the year 2020 there may be as many as 35 million vehicles on British roads.***
>
> ***Alternative forms of transport should be encouraged and international laws introduced to control car ownership and use.***
>
> ***To what extent do you agree or disagree?***

Give reasons for your answer and include any relevant examples from your knowledge or experience.

Write at least 250 words.

2 MONT-TI-22

Currently, the main global issue is global warming. We need to control our energy use and make smarter choices when it comes to the planet. Regulating transportation, is one of the ways we could reduce our energy consumptions. Indeed, the use of individual cars is now part of our society's idea of a "nice life", as it was so heavily structured throughout the 20th century. Knowing what we know now about the greenhouse effect and global warming it would be best to encourage the development of other forms of transportation. Public transportation, should be expended to more rural areas and spread usage. Here, making individual vehicles less necessary for everyday life. This change need to come from governments, as it unfortunately by actions affecting everyone which can't be just the individuals' responsability. That is why it is so important to vote for people whose priorities are fighting global warming and developing a sustainable way of life. Even though global warming is everyone's responsibility as it affects us all, people in power with economical means need to make these choices to overcome easier. Again, it is hard to ask people with very few means of transportation to reduce their use of vehicles if no other options is presented to them. Still, when it comes to issues, governments do have a role to play is influencing people, discouraging them from purchasing new vehicles and opting for other options such as public transportation, or cleaner vehicles.

To conclude, I completely agree with the fact that national and international governments and legislations have an important part to play in the reduction of car ownership and use and global warming in general.

Appendix F:

Sample Responses of Vocabulary Tasks and IELTS Writing (Chapter Six) (Testing Time One)

Lex30 (Meara & Fitzpatrick, 2000)

Lex30

(Meara & Fitzpatrick, 2000)

Name (in English):..... Code:.....

Date: ...../...../

Time: 15 minutes
Instruction: Write down the first four (English) words you think of when you read each word in the list.

1.	attack	accident	volleyball	shock	broke
2.	board	tree	skiing	square	hard
3.	close	shop	end	night	open
4.	cloth	shirt	shopping	buy	look
5.	dig	shovel	ground	hand	clog
6.	dirty	room	clean	wash	sorb
7.	disease	hospital	doctor	medicine	scared
8.	experience	job	study	advantage	important
9.	fruit	tasty	delicious	apple	bananas
10.	furniture	sofa	bed	big	lamp
11.	habit	dairy	anytime	good	bad bad
12.	hold	hand	hug	bag	keep
13.	hope	happy	kind	die	dangerous
14.	kick	leg	soccer	socks	shoes
15.	map	place	find	lost	Google
16.	obey	teacher	parents	boss	prosen president
17.	pot	water	boil	hot	plug
18.	potato	vegetable	brown	tasty	ground
19.	real	dream	money	strict	sleep
20.	rest	vacation	time	relax	coffee
21.	rice	white	Japan	curry	eat
22.	science	class	school	experiment	difficult
23.	seat	sit	theater	restaurant	chair
24.	spell	English	wrong	test	write
25.	substance	content	mount	gram	real
26.	stupid	foolish	crime	embarrassing	bad
27.	television	show	watch	drama	comedy
28.	tooth	important	white	toothbrush	bite
29.	trade	buy	change	business	item
30.	window	open	close	clear	glass

G_Lex (Fitzpatrick & Clenton, 2017)

G-Lex
(Fitzpatrick & Clenton 2017)

Name (in English):..... Code:.....-.....

Date: ...../...../

Time: 15 minutes
Instruction: Write down five different words that might fit into each gap. The gaps are suitable for nouns, adjectives, and verbs in equal measure (eight sentences each).

1. She loved to _____ over the phone.	take	hold	call	hand	give
2. When I feel sad I always go to the_____.	sea	hometown	gym	park	bed
3. They think car-racing is_____.	interesting	exciting	boring	great	dangerous
4. His colleague wanted to _____ the report.	start	write	read	put	take
5. My favourite __ ___ is football.	sport	hobby	class	club	exercise
6. She looked _____ when she saw her friends.	happy	excited	sad	bad	angry
7. He couldn't _____ the car.	~~dr~~drive	get	have	clean	repair
8. With a fire in my house I would save my ____.	money	bags	clothes	card	memorise
9. Many people feel ____about the environment.	good	bad	uncomfortable	comfortable	complain
10. The parents _____ the children.	love	angry	grow	like	ignore
11. He was happy with his _____.	dog	parents	friends	girlfriend	money
12. He didn't think her teacher was _____ at all.	singer	actor	angry	dancer	married
13. She always wanted to _____ after a busy day at work.	drink	exercise	dance	sing	drive
14. She sent _____to her mother.	email	letter	money	thing	present
15. The weather looked ____ before the game.	good	bad	nice	cloudy	cold
16. He wanted to _____ the letter.	write	send	read	get	give
17. She was excited about _____.	it	game	snow	show	movie
18. The girls thought the rock concert was____.	exciting	interesting	good	great	awesome
19. He took the chance to _____ the president.	be	see	meet	talk	hug
20. He gave his boss _____.	documents	books	call	messages	sweets
21. At the funeral the family felt ___.	sad	bad	crying	dark	silent
22. He always _____his breakfast.	eat	hers	forget	leave	
23. She put the food in the _____.	bag	refreiglater	shelf	boys	mouth
24. She was always ____ to those who needed help.	kind	smile	angry	strict	helpful

Productive Vocabulary Level Test (the PVLT; Laufer & Nation, 1999)

Productive Vocabulary Level Test
(PVLT; Laufer and Nation, 1999)

Name (in English):..... Code:.....

Date: ...../...../

Time: 25 minutes
Instruction: Complete the underlined words. The example has been done for you.
He was riding a bicycle.

The 2000-word level

1. I am glad we had this opportunity to talk.
2. There are a dozen eggs in the basket.
3. Every working person must pay income tax.
4. The pirates buried the treat on a desert island.
5. Her beauty and charm had a powerful effect on men.
6. Lack of rain lead to a shortage of water in the city.
7. He takes cream and sugar in his coffee.
8. The rich man died and left all his we_____ to his son.
9. Pup_____ must hand in their papers by the end of the week.
10. This sweater is too tight. It needs to be stret_____.
11. Ann introduce her boyfriend to her mother.
12. Teenagers often admire and worship pop singers.
13. If you blow up that balloon anymore it will burn.
14. In order to be accepted into the university, he had to improve his grades.
15. The telegram was delivary to ours after it had been sent.
16. The differences were so slightly that they went unnoticed.
17. The dress you are wearing is lovely.
18. He wasn't very popular when he was a teenager, but he has many friends now.

The 3000-word level

1. He has a successful carrer as a lawyer.
2. The thieves threw accident in his face and made him blind.
3. To improve the country's economy, the government decided on economic refer.
4. She wore a beautiful green go_____ to the ball.
5. The government tried to protect the country's industry by reducing the improvement cheap goods.
6. The children's games were funny at first, but finally got on the parents' ner_____.
7. The lawyer gave some wise coun_____ to his clients.
8. Many people in England mow the la_____ of their houses on Sunday morning.
9. The farmer sales the eggs that his he_____ lays.
10. Sudden noises at night sca_____ me a lot.
11. France was proc_____ a republic in the 18th century.
12. Many people are injured in road accidents every year.
13. Suddenly, he was thrught into the dark room.
14. He perc_____ a light at the end of the tunnel.
15. Children are not independent. They are attack to their parents.
16. She showed off her sle_____ figure in a long narrow dress.
17. She has been changing partners often because she cannot have a staraigt relationship with one person.
18. You must wear a bathing suit on a public beach. You're not allowed to be na_____.

The 5000-word level

1. Soldiers usually swear an oa _____ of loyalty to their country.
2. The voter placed the ball_____ in the box.
3. They keep their valuables in a vau_____ at the bank.
4. A bird perched at the window led er .
5. The kitten is playing with a ball of ya_____.
6. The thieves have forced an ent er into the building.
7. The small hill was really a burial moun tain .
8. We decided to celebrate new year's e_____ together.
9. The solider was asked to choose between infinitary and cav_____.
10. This is a complex problem which is difficult to compr_____.
11. The angry crowd sho w the prisoner as he was leaving the court.
12. Don't pay attention to this rude remark. Just ign ore it.
13. The management held a secret meeting. The issues discussed were not disc ussed to the workers.
14. We could hear the sergeant bel_____ commands to the troops.
15. The boss got angry with the secretary and it took a lot of tact to soo_____ him.
16. We do not have adeq_____ information to make a decision.
17. She is not a child, but a mat_____ woman. She can make her own decisions.
18. The prisoner was put in soli_____ confinement.

The University Word List Level

1. There has been a recent tr_____ among prosperous families towards a smaller number of children.
2. The ar_____ of his office is 25 square meters.
3. Phil_____ examines the meaning of life.
4. According to the communist doc ument workers should rule the world.
5. Spending many years together deepened their inti_____.
6. He usually read the sport sec tions of the newspaper first.
7. Because of the doctors' strike the cli_____ is closed today.
8. There are several misprints on each page of this te xt .
9. The suspect had both opportunity and mot_____ to commit the murder.
10. They insp ect all products before sending them out to stores.
11. A considerable amount of evidence was accum_____ during the investigation.
12. The victim's shirt was satu_____with blood.
13. He is irresponsible. You can not re_____ on him for help.
14. It's impossible to eva_____ these results without knowing about the research methods that were used.
15. He finally att end. a position of power in the company.
16. The story tells us about a crime and subs cri be punishment.
17. In a hom eroom class all students are of a similar proficiency.
18. The urge to survive is inh abit in all creatures.

The 10000-word level

1. The baby is wet. Her dia_____ needs changing.
2. The prisoner was released on par_____.
3. Second year University students in the US are called soph_____.
4. Her favorite flowers were or_____.
5. The insect causes damage to plants by its toxic sec_____.
6. The evac_____ of the building saved many lives.
7. For many people, wealth is a prospect of unimaginable felic_____.
8. She found herself in a pred_____ without any hope for a solution.
9. The deac_____ helped with the care of the poor of the parish.
10. The hurricane whi_____ along the coast.
11. Some coal was still smol_____ among the aches.
12. The dead bodies were muti_____ beyond recognition.
13. She was siting on a balcony and bas_____ in the sun.
14. For years waves of invaders pill_____ towns along the coast.
15. The rescue attempt could not proceed quickly. It was imp_____ by bad weather.
16. I wouldn't hire him. He is unmotivated and indo_____.
17. Computers have made typewriters old-fashioned and obs_____.
18. Watch out for his wil_____ tricks.

IELTS Writing Topic and Its Sample Response (Chapter Six) (Testing Time One)

You should spend about 40 minutes on this task.

Present a written argument or case to an educated reader with no specialist knowledge of the following topic.

Some people believe that teaching children at home is best for a child's development while others think that it is important for children to go to school. Discuss the advantages of both methods and give your own opinion.

Give reasons for your answer and include any relevant examples from your own knowledge or experience. Write at least 250 words.

First, the advantage of teaching children at home is convenience and safety. For example, teaching at home does not need rules, wamy, school uniforms, lunch boxes, and so on, So their parents, it is much better I think. But it also is difficult for some parents to teach how to solve a question, the reason why it is. And it is harol for children to get a lot of ~~experiments~~ experience and get a good relationship too. Second., the advantage of going to school is teaching by expert, getting experience, learning how to make a relationship. For example, children are easier to understand how to correct and the reason why it is. They also can get knowledge through a lot of happens and events. And they can hang out with a new friend and make an esciting memory. But going to school have possible about happening trouble between students. And if a student does not have enough money, it is hard to go there, because they need to buy a school uniform, textbook and so on. Finally, in my opinion, I would like to choose to go to school.. By going to school we can meet a lot of people and get ~~many~~ many opportunities to have some dreams or notice what I like and I want to do. And now in life at school, I have good friends and dreams.

You should spend about 40 minutes on this task.

Present a written argument or case to an educated reader with no specialist knowledge of the following topic.

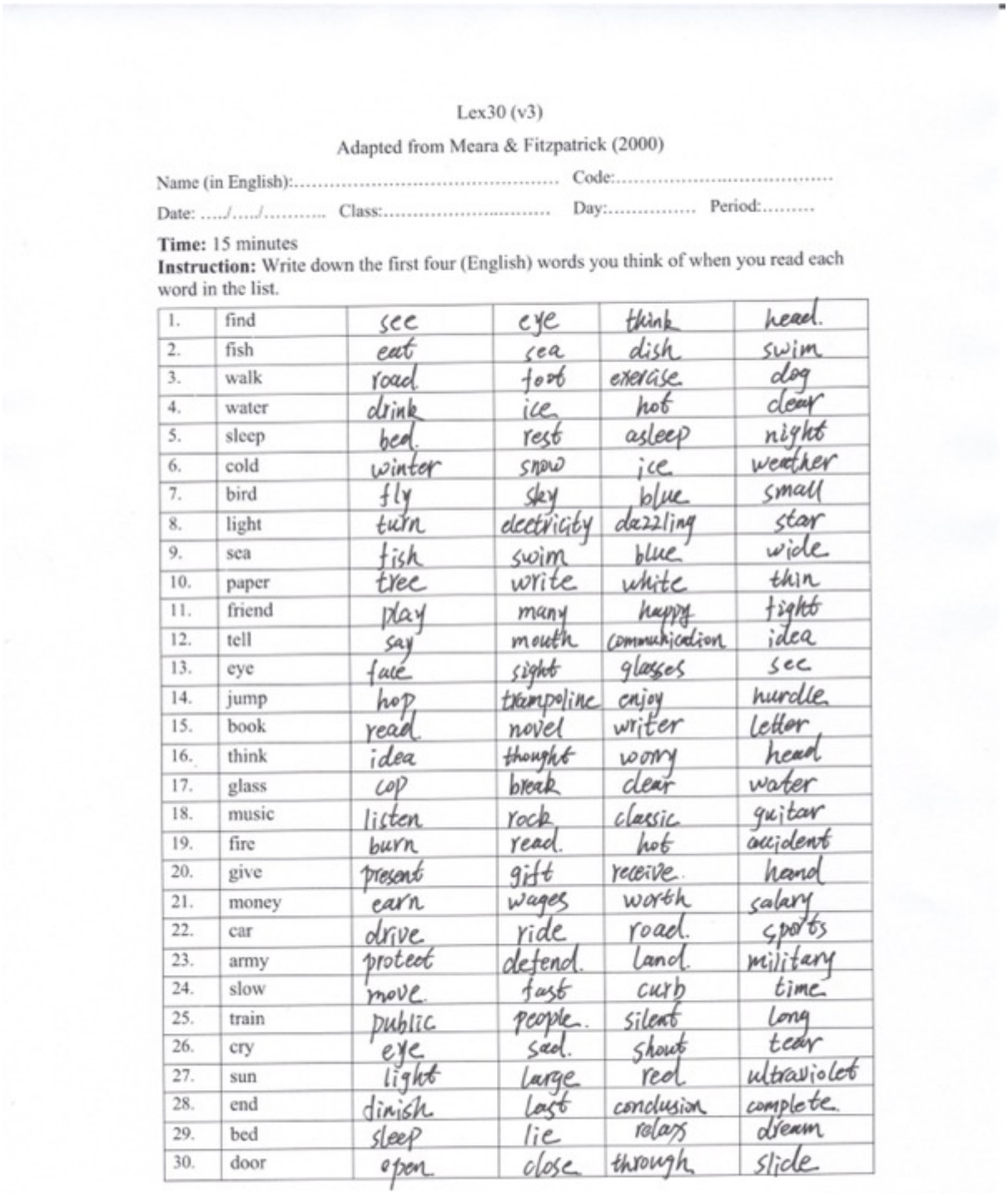
At the present time, the population of some countries includes a relatively large number of young adults, compared with the number of older people. Do the advantages of this situation outweigh the disadvantages?

Write at least 250 words.

No, it does not, I think that including a large number of young adults is not better. But of cause, there is the advantage that young people have a lot of possible and future. So this is a good point because young people make and more develope life, society, and the world in the future. And it is bad to say but older people just have knowledge and experience. I know that it is a very important thing because young people are told that by them. Then I start to talk about the disadvantage First, and this is the most serious issue. It is hard for young actua adults to find and get a job. Because the population of young people is large. Recently people who do not work are many. I think this is the reason why the situation is caused by. In addition to that issue, in the future, it is said that a lot of jobs are disappointed because the technology of AI is more developing than now. Second, there is possible about the culture and tradition of the country are becoming old easily. In Japan, there is some gap between young adults and older people about language, knowledge, thinking a thing. So other countries also happen the same thing. Maybe this is said that change the era, but remember and kind tradition is important. Third, when young adults are older and if the population of the next generation is not large, it is hard for them to support the older people. For example, tax and care for older people. Japan has this issue. Through these, I think the advantage of a relatively large number of young adults, compared with the number of older people do not outweigh.

Sample Response of Vocabulary Tasks and IELTS Writing (Chapter Six) (Testing Time Two)

Lex30 (Meara & Fitzpatrick, 2000)



Lex30 (v3)

Adapted from Meara & Fitzpatrick (2000)

Name (in English):..... Code:.....

Date: ...../...../..... Class:..... Day:..... Period:.....

Time: 15 minutes

Instruction: Write down the first four (English) words you think of when you read each word in the list.

1.	find	see	eye	think	head.
2.	fish	eat	sea	dish	swim
3.	walk	road	foot	exercise	dog
4.	water	drink	ice	hot	clear
5.	sleep	bed	rest	asleep	night
6.	cold	winter	snow	ice	weather
7.	bird	fly	sky	blue	small
8.	light	turn	electricity	dazzling	star
9.	sea	fish	swim	blue	wide
10.	paper	tree	write	white	thin
11.	friend	play	many	happy	fight
12.	tell	say	mouth	communication	idea
13.	eye	face	sight	glasses	see
14.	jump	hop	trampoline	enjoy	hurdle
15.	book	read	novel	writer	letter
16.	think	idea	thought	worry	head
17.	glass	cup	break	clear	water
18.	music	listen	rock	classic	guitar
19.	fire	burn	read	hot	accident
20.	give	present	gift	receive	hand
21.	money	earn	wages	worth	salary
22.	car	drive	ride	road	sports
23.	army	protect	defend	land	military
24.	slow	move	fast	curb	time
25.	train	public	people	silent	long
26.	cry	eye	sad	shout	tear
27.	sun	light	large	red	ultraviolet
28.	end	finish	last	conclusion	complete
29.	bed	sleep	lie	relax	dream
30.	door	open	close	through	slide

G_Lex (Fitzpatrick & Clenton, 2017)

G-Lex (v3)
Adapted from Fitzpatrick & Clenton (2017)

Name (in English):..... Code:.....

Date: ...../...../..... Class:..... Day:..... Period:.....

Time: 15 minutes
Instruction: Write down five different words that might fit into each gap. The gaps are suitable for nouns, adjectives, and verbs in equal measure (eight sentences each).

1. He tried to _____ during his summer vacation.	exercise	study	trip	read	run
2. When I feel angry I always go to the_____.	bed	walk	room	bath	hometown
3. They think football is_____.	fun	difficult	hard	sport	nice
4. She wanted to _____ the project.	join	quiet	do	inform	start
5. My best _____is orange.	food	color	fruit	thing	diet
6. She felt _____ when she received her test score.	happy	good	angry	disappointed	crying
7. She couldn't _____ the house.	go	build	visit	see	invite
8. She should include more ____ in her next report.	content	reason	thought	fact	basis
9. My friends feel ____ about my new car.	good	nice	better	bad	nothing
10. The government _____ the people.	inform	say	show	ask	please
11. He was surprised about his _____.	style	face	voice	car	character
12. He thought his parents were _____.	kind	nice	good	happy	angry
13. She wanted to _____ her life.	improve	enjoy	change	relax	success
14. She sent _____to her boss.	e-mail	letter	book	paper	present
15. We looked ____ before the game.	him	her	watch	drama	player
16. He wanted to _____ the message.	receive	get	leave	read	say
17. She was happy about her _____.	life	class	score	job	husband
18. They thought the basketball game was____.	nice	good	so so	bad	perfect
19. He tried to _____ his teacher.	ask	be	talk	change	memorize
20. She gave her mother _____.	cake	bag	book	thanks	message
21. At the graduation party, the family felt ___.	crying	moving	touching	nice	good
22. She always _____her bag.	hold	forget	have	put	look
23. Last night, I had my worst _____.	time	day	game	food	life
24. They are_____ players.	good	soccer	baseball	tennis	kids

Productive Vocabulary Level Test (the PVLT; Laufer & Nation, 1999)

Productive Vocabulary Level Test (**v2**)
(PVLT; Laufer and Nation, 1999)

Instruction: Complete the underlined words. The example has been done for you.
He was riding a bicycle.

The 2000-word level

1. It is the de_____ that counts, not the thought.
2. Plants receive water from the soil through their ro_____.
3. The nurse was helping the doctor in the operation room.
4. Since he is unskilled, he earns low wages.
5. This year long skirts are fashionable again.
6. Laws are based upon the principle of justice.
7. He is walking on the tile of his toes.
8. The mechanic had to replace the motor of the car.
9. There is a copy of the original report in the file.
10. They had to climb a steep mountain to reach the cabin.
11. The doctor exists the patient thoroughly.
12. The house was su_____ by a big garden.
13. The railway continue London with its suburbs.
14. She wan_____ aimlessly in the street.
15. The organisers li_____ the number of participants to fifty.
16. This work is not up to your usual standard.
17. They sat down to eat even though they were not hungry.
18. You must have been very br_____ to participate in such a dangerous operation.

The 3000-word level

1. I live in a small apartment on the second floor.
2. The profile of failing the test scared him.
3. Before writing the final version, the student wrote several drama.
4. It was a cold day. There was a chill in the air.
5. The cart is pulled by an older.
6. Anthropologists study the struction of ancient societies.
7. After two years in the Army, he received the rank of lieutenant
8. The statue is made of marterial.
9. Some aristocrats believed that blue blood flowed through their ve_____.
10. The secretary assignment the boss in organizing the course.
11. His beard was too long. He decided to treat it.
12. People were whir_____ round on the dance floor.
13. He was on his knees, please for mercy.
14. You'll sn_____ that branch if you bend it too far.
15. I won't tell anybody. My lips are sea_____.
16. Crying is a nor_____ response to pain.
17. The Emperor of China was the supr_____ ruler of his country.
18. You must be aware that very few jobs are available.

The 5000-word level

1. Some people find it difficult to become independent. Instead they prefer to be tied to their mother's ap_____ strings.

1/3

2. After finishing his degree, he entered upon a new ph_____ in his career.
3. The workmen cleaned up the meeting room before they left.
4. On Sunday, in his last se_____ in Church, the priest spoke against child abuse.
5. I saw them sitting on st_____ at the bar drinking beer.
6. Her favorite musical instrument was a trumpet.
7. The building is heated by a modern heating appa_____.
8. He received many com_____ on his dancing skill.
9. People manage to buy houses by raising a mor_____ from a bank.
10. At the bottom of a blackboard there is a le_____ for chalk.
11. After falling off his bicycle, the boy was covered with bru_____.
12. The child was holding a doll in her arms and hugging it.
13. We'll have to be inventive and de_____ a scheme for earning more money.
14. The picture looks nice; the colours bloom really well.
15. Nuts and vegetables are considered who_____ food.
16. The garden was full of fra_____ flowers.
17. Many people feel depressed and gl_____ about the future of the mankind.
18. He is so depressed that he is cont_____ suicide.

The University Word List Level

1. I've had my eyes tested and the optician says my vi_____ is good.
2. The anom_____ of his position is that he is the chairman of the committee, but isn't allowed to vote.
3. In their geography class, the children are doing a special pro_____ on North America.
4. In a free country, people can apply for any job. They should not be discriminated against on the basis of colour, age, or sex.
5. A true dem_____ should ensure equal rights and opportunities for all citizens.
6. The drug was introduced after medical rescue indisputably proved its effectiveness.
7. These courses should be taken in sequence, not simultaneously.
8. Despite his physical condition, his int_____ was unaffected.
9. Governments often cut budgets in times of financial critical.
10. The job offer sounded interesting at first. But when he realised what it would involve, his excitement subs_____ gradually.
11. Research indeX that men find it easier to give up smoking than women.
12. In a lecture, most of the talking is done by the lecturer. In a seminar, students are expected to participate in the discussion.
13. The airport is far away. If you want to ensure that you catch your plane, you have to leave early.
14. It's difficult to ass_____ a person's true knowledge by one or two tests.
15. The new manager's job was to res_____ the company to its former profitability.
16. Even though the student didn't do well on the midterm exam, he got the highest mark on the final.
17. His decision to leave home was not well thought out. It was not based on rat_____ considerations.
18. The challenging job required a young, successful and dynamic candidate.

The 10000-word level

1. The new vic_____ was appointed by the bishop.
2. If your lips are sore, try lip sal_____, not medicine.
3. Much to his chag_____, he was not offered the job.
4. The actors exchanged ban_____ with reporters.
5. She wanted to marry nobility: a duke, a baron, or at least a visual.
6. The floor in the ballroom was a mos_____ of pastel colours.
7. She has contributed a lot of money to various charities. She is known for her generosity and benefib.
8. This is an unusual singer with a range of three oct_____.
9. A throwing controls the flow of gas into an engine.
10. Anyone found loo_____ bombed houses and shops will be severly punished.
11. The crowd soon dispose when the police arrived.
12. The wounded man squi_____ on the floor in agony.
13. The dog crin_____ when it saw the snake.
14. He immediate himself in a hot bubbly bath forgetting all his troubles for a moment.
15. The approaching storm stam_____ the cattle into running wildly.
16. The problem is beginning to assume mam_____ proportions.
17. His vind_____ behaviour towards the thief was understandable.
18. He was arrested for illi_____ trading in drugs.

IELTS Writing Topic and Its Sample Response (Chapter Six) (Testing Time Two)

You should spend about 40 minutes on this task.

Present a written argument or case to an educated reader with no specialist knowledge of the following topic.

Nowadays, adults do little exercise. Some people believe that the best way to address this issue is by covering great sports events such as the Olympics on television. Others think that it is more beneficial to take other measures. What is your opinion?

Give reasons for your answer and include any relevant examples from your own knowledge or experience. Write at least 250 words.

I do not think that the best way to address the issue is by covering great sports events such as the Olympics on television. Certainly, I think that not a few people start exercising under the influence of the Olympics. However, I guess that many people just enjoy watching the players. Not only that, children are more influenced and start exercising than adults. First, many adults find it difficult to spend time exercising because they are working. In Japan, people who work have a lot of overtime, and when they go home they are so tired and can not exercise. In addition, they have relationships with co-workers and seniors, and even if they are full of child-rearing and housework. Second, it costs a lot of to start exercising. For example, if you start playing tennis or soccer, you need a race or shoes. Also, even if you go to the gym, you have to pay the membership fee. So the hurdle is high if you are not sure that you will continue the exercise. I came up with two ideas to solve the lack of exercise for adults. The first is to walk even a little when commuting. People who commute by train or bus walk one station. people who use a car either change to a bicycle or park a bit little farther than usual and walk. I think it makes them feel better to do it before, not after work that they are tired of. The second is to buy goods that can be easily done at home to eliminate little exercise This is cheaper than starting a sport or going to the gym. And anyone can it because they only have to make a little freetime at home. The most important thing is how easy it is and the low hurdless to get started.

You should spend about 40 minutes on this task.

Present a written argument or case to an educated reader with no specialist knowledge of the following topic.

People believe that using mobile phones and computers to communicate makes us lose the ability to communicate with each other face to face. To what extent do you agree or disagree?

Give reasons for your answer and include any relevant examples from your own knowledge or experience. Write at least 250 words.

I partly agree that using mobile phones and computers to communication makes us lose the ability to communicate with each other face to face. We spend a lot of time on mobile phones and computers for entertainment such as watching videos and playing games. By doing this, we spend more time on our own and less time and opportunity to communicate with friends and family. Furthermore, SNS such as Twitter and Instagram are developping. But these are mainly to let an unspecified number of people know their lives and thoughts. This is one-side relationship, not a communication that should be reciprocal. As proof of that, recently, I often get sick and commit suicide based on SNS posts, I think this is because it is one-side. And you can think of it as having poor communication skill. On the other hand, mobile phones and computers can also communicate face-to-face. For example, LINE and ZOOM can use viedo. This allows us to see each other's face and speak directly. With the spread of corona-virus infection, it may be clear that communication can be achieved using these tools. In conclusion, using mobile phones and computers make us lose the ability to communicate with each other face-to-face. ~~However~~ However, I think the most important thing is whether the person uses it as an entertainment or one-side tool, or for other purpose.