Investigating Factors Affecting Lexical Diversity Measure Predictions of Writing and Speaking Proficiency: Word-Counting Criteria, L1 Background, Language Proficiency, and Text Length

by

Thwin Myint Myint Maw

B.A., Meiktila University, 2005

M.A., Meiktila University, 2009

M.Ed., University of Fukui, 2020

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

Graduate School of Humanities and Social Sciences

(Integrated Arts and Human Sciences Program)

HIROSHIMA UNIVERSITY

March 2023

Acknowledgements

Foremost, my heartfelt gratitude goes to my main academic supervisor, Dr. Jon Clenton for his invaluable guidance and continuous support throughout my doctoral program. His encouragement and patience inspired and truly helped me finish this study. I am also very much grateful to the sub-committee members, Dr. Noriko Yamane, and Dr. Masahiro Shinya for their helpful advice and recommendations on this dissertation. I reserve special thanks for help from Professor Simon Fraser.

I would also like to say my special thanks to Dr. George Higginbotham, whose great help throughout my PhD journey. I faced challenges of conducting experiments during Covid-19. But he kindly helped me overcome this barrier by conducting the experiments with his students at The Language Center: Queen Mary University of London and also providing feedback on my paper and thesis writing. I also owe thanks to the staff and students for their kind support and active participation in the experiments.

My sincere appreciation also goes to Dylan Jones, who assisted me by proofreading and providing the valuable comments on my paper and the entire thesis despite his tight schedule.

Finally, I am incredibly grateful to everyone who supported me personally, academically, or professionally during my PhD journey.

Summary

Lexical diversity (i.e., the different words) used in a written or spoken text is crucial in estimating L2 language proficiency. Various LD measures have been developed for vocabulary and language assessment. Because of the text length sensitivity of basic measures (*Types*; simple count of every word that occur once, and *Type-Token-Ratio*; the proportion of different words to total words), sophisticated measures with more complicated quantifications have been formulated. Previous research has validated existing LD measures' applicability in the L2 context and has found that LD measures can be reliable L2 general, writing, and speaking proficiency indicators.

Researchers (e.g., Treffers-Daller et al., 2018; Yu, 2010; Zenker and Kyle, 2021) have identified important factors that can influence the accuracy of LD measures in predicting these wider L2 proficiencies, namely: the analysis units used, L1 background, language proficiency, and text length. However, previous validation studies have been lacking in address and controlling and incorporating these four factors into L2 lexical diversity assessment because the studies have considered only one or two of these factors.

The current dissertation, therefore, addresses this important gap in LD research. The dissertation investigates whether LD measures predict inter- and intra-group writing variability under a controlled text length (200 words) and for a specific L1 background (Chinese). It also examines the extent to which LD measures predict speaking proficiency based on using different constant spoken-sample text lengths (200 to 450 words). It examines the extent to which LD measures predict writing and speaking using different-word counting techniques, focusing on the utility of the

flemma count (a base word and its inflections under different word classes as the same types).

The dissertation comprises four experiments, the partial replications of Treffers-Daller et al.'s (2018) study. The first experiment was based on an entire population (N = 194 L2 English writers from mixed L1 backgrounds). It investigated the extent to which LD measures could discriminate between the three IELTS-based writing proficiency levels (6.5, 7, 7.5) under a controlled text length based with the different analysis units. The second experiment controlled the L1 background and so examined the extent to which LD measures predict the writing proficiency of an L1 Chinese L2 English learner group (N = 105).

The third experiment controlled both L1 background and language (writing) proficiency. It explored the extent to which LD measures to predict writing proficiency of L1 Chinese L2 English learners (N = 103) based on different writing proficiency levels (6.5, 7, 7.5). The fourth experiment analyzed the different participants (55 L2 English speakers from various L1 backgrounds). It examined whether LD measures were predictive of the IELTS-based speaking proficiency levels (6.5, 7, 7.5) based on the different analysis units and text lengths. It followed similar procedures to the first writing experiment to gain greater comparability of the findings of the LD measure predictions of two different language modes (L2 writing and speaking).

Overall, the four experimental studies' findings indicate that different analysis units influenced LD measure predictions of L2 language proficiency. Furthermore, LD measures were stronger L2 writing predictors than speaking predictors, and LD measures required longer constant text length for speaking than writing to achieve accurate predictability. This thesis concludes that LD measure predictions of L2

language proficiency is dependent on these four factors, so future LD research should consider and control them carefully.

Table of Contents

Page

List of Tables

List of Figures

Figure **Page**

Chapter 1

Introduction

1.1 Overview

This dissertation aims to validate lexical diversity (LD) measures' ability to predict L2 language (writing and speaking) proficiency by controlling four factors (word-counting criteria, L1 background, language proficiency, and text length) that can influence their accuracy. These factors and their effects have not been sufficiently or systematically addressed in previous studies on the assessment of L2 lexical diversity.

This introductory chapter comprises four sections. Section 1.1 presents an overview of this introduction. Section 1.2 briefly explains what lexical diversity means, notes LD measures' sensitivity to text length variation, and discusses research attempts to overcome this text length variation issue. Section 1.3 describes the existing research evidence of LD measures' ability to predict L2 language proficiency, highlighting several factors that might influence the predictive validity of the LD measures. Section 1.4 identifies the need to control the four main influential factors in order to validate LD measures' applicability in L2 lexical diversity assessment.

1.2 Lexical diversity (LD) and its measurement

Lexical diversity, also known variously as lexical variation, vocabulary range, and vocabulary richness, refers to the variety of words used in a spoken or written sample (Read, 2000). Simply put, it is "the proportion of words in a language sample that are not repetitions of words already encountered" (Jarvis, 2013, p. 88). Lexical diversity has been useful in predicting general L2 English language proficiency. The

most basic LD quantification method is simply to count the different numbers of unique words found in a text. However, this method is very sensitive to text length variation since longer texts are likely to include more repeated words, resulting in lower lexical diversification.

To overcome this text sample size issue, both basic LD measures using mathematical transformations (*TTR, Guiraud's Index*) and more sophisticated and stable measures (*D, MTLD, HD-D*) have been developed. However, these existing LD measures remain sensitive to text length. Thus, an alternative means to counter this issue is to use standardized text length (Duran et al., 2004; Treffers-Daller, 2013; Treffers-Daller et al., 2018) for reliable LD score comparability.

1.3 Factors influencing LD measures' applicability to L2 language proficiency assessment

With the recognition of the crucial role that vocabulary range can play in estimating L2 language proficiency, the usability of both basic and sophisticated LD measures has long been validated. Studies examining spoken and written LD have shown that LD measures are important indicators of speaking, writing, and general language proficiency (Crossley & McNamara, 2013; Gonzalez, 2017; McCarthy & Jarvis, 2010; Nasseri & Thompson, 2021; Read & Nation, 2006; Treffers-Daller et al., 2018; Vögelin et al., 2019; Wang, 2014; Wu et al., 2019; Yu, 2010).

However, researchers also acknowledge that there are multiple factors affecting LD measure usability in the L2 context. Four widely acknowledged factors are the word counting criteria, L1 background, language proficiency, and text sample size.

First, as an obstinately ongoing issue, LD measures remain influenced by text length to varying degrees and have shown different levels of stability at different text

lengths (Zenker & Kyle, 2021). Second, the written or spoken diverse vocabulary use of L2 learners from varying L1 backgrounds has been shown to be different (Clavel-Arroitia & Pennock-Speck, 2021; Yu, 2010). Yu (2010), for example, found no significant lexical diversity differences between L1 Filipinos and L1 Chinese L2 English learners. Clavel-Arroitia and Pennock-Speck (2021) also found that L1 Spanish L2 English learners had higher written and spoken lexical diversity than L1 Japanese L2 English learners. Thus, the L1 background appears to be a potential factor affecting LD measures' predictive validity.

The third factor is an issue in L2 vocabulary assessment that has become controversial in recent times: the suitability of different word-counting units for L2 learners who might have different word knowledge levels, as schematized by Bauer and Nation (1993). Recent studies (Brown, 2018; Brown et al., 2020; McLean, 2017; Stoeckel et al., 2020) have proposed using two smaller word-counting units over a traditional word-family count (a word's both inflected and derived forms as the same type) for L2 learners with limited derivational knowledge. These two alternative units demanding only learners' inflectional knowledge are the flemma (inflections under different word classes as the same word type) and the lemma (inflections under different word classes as different word types); using flemma and lemma counts can thus produce different evaluations of lexical diversity.

We can see the wide extent to which different word-counting criteria impact studies in a range of different studies, including LD studies, which have paid little attention to the analysis units and thus have somewhat randomly employed different units. However, a recent contributive study to the current LD literature, Treffers-Daller et al.'s (2018) study, raised awareness of the cruciality of careful analysis unit choice in L2 lexical diversity assessment. The authors proved that the ability of LD

measures to predict L2 language proficiency varied depending on the word-counting methods used, highlighting the need to control the effects of the word-counting technique on LD measures and scores.

Fourth, the level of L2 language proficiency might be another potential factor influencing the extent to which LD measures predict L2 proficiency. L2 learners at different proficiency levels might have different lexical knowledge. For instance, because of their larger vocabulary size, higher-proficiency learners' texts might be more lexically diverse than lower-proficiency learners' texts. Also, learners might use inflections and derivations differently in their written or spoken products. Leontjev et al. (2016) found that L2 learners' derivational knowledge increased with higher writing proficiency. Therefore, LD measures' discrimination between within-level variations of L2 writers or speakers might vary based on their proficiency levels. However, the extent to which LD measures predict intra-group language variability has received relatively little attention in L2 lexical diversity assessment so far.

1.4 The need to address four influential factors (word-counting criteria, L1 background, language proficiency, text length) for greater validity of LD measures' L2 language predictions

Research into LD measures' usability in L2 language assessment has amply proved different aspects of their validity, particularly their predictive validity. Previous LD researchers have confirmed that LD measures are reliable predictors of L2 writing, speaking and general language proficiency (Gonzalez, 2017; Jarvis, 2013; Jarvis, 2017; McCarthy & Jarvis, 2013; Nasseri & Thompson, 2021; Read & Nation, 2006; Treffers-Daller et al., 2018; Yu, 2010; Zenker & Kyle, 2021; Zhang & Daller, 2019). However, these researchers' findings were drawn from studies addressing only one or two of the above four factors.

First, the analysis unit choice has received relatively little attention in L2 lexical diversity assessment, and so most previous studies have applied various analysis units that demand L2 learners' different inflectional and derivational knowledge. Most LD studies have simply counted all the different words used in a text as different types (simple count). Yu (2010) intentionally used a simple count because of the initial analysis findings of the participants' use of few inflections. Jarvis (2013) and Nasseri and Thompson (2021) considered a word and its inflections under the same word class as the same type (ie., a lemma count). Recently, Treffers-Daller et al. (2018) conducted a wider examination, comparing the three different analysis units of simple, lemma, and word-family (a word and both its inflections and derivations as the same type) counts.

Despite LD studies examining the suitability of simple, lemma, and word-family counts in L2 lexical diversity assessment, no single empirical study has, to my knowledge, yet reported on LD measure predictions of L2 language proficiency based on a flemma count (a word and its inflections regardless of word class). The current study, therefore, represents the first study to investigate the influence of a flemma count on LD measure predictions compared to simple and lemma counts.

Second, despite the acknowledgement of the influence of L1 background influence on LD measures and scores (Yu, 2010), most LD studies' findings and conclusions about LD measure predictions were based on investigations of L2 learners from diverse L1 backgrounds (Read & Nation, 2006; Treffers-Daller et al., 2018; Zenker & Kyle, 2021), and so were uniformly applied to mixed L1 backgrounds, which may not be appropriate.

However, L2 learners from different L1 backgrounds might make varying use of diverse vocabulary, and that might affect LD measures' predictive ability. For

instance, Clavel-Arroitia and Pennock-Speck (2021) identified higher lexical diversity knowledge of L1 Spanish L2 English learners than L1 Japanese L2 English learners. Yu (2010) found that an LD measure (*D*) was significantly predictive of L2 proficiency of the entire population for L2 learners from multiple L1 backgrounds, but it could not differentiate between L1 Chinese and L1 Filipino L2 English learners. Given the existing empirical evidence of the effects of different L1 backgrounds on LD measures and scores, there is a need for more research into the dependence of LD measure predictions of L2 language proficiency on L1 background.

Third, despite rich evidence of LD measure predictions of inter-group language (e.g., writing) variations, LD measures' discernment of intra-group variability is still unclear. Because of the widely diverse language proficiency of L2 learners, it is also necessary to explore the extent to which LD measures are useful in discriminating between within-level language differences in the L2 context. Therefore, this thesis investigates the effects of language proficiency on LD measure predictions.

Fourth, although LD measures remain sensitive to text sample size, most previous LD studies have validated LD measure predictions by using written or spoken texts of differing lengths. Thus, Treffers-Daller et al. (2018) carefully set a constant text length for their study because comparing LD scores from texts of varying length seemed inappropriate. Similarly, to support LD measure predictions under controlled text length, the current study addresses the text length influence on LD measure predictions of L2 language (writing and speaking) proficiency.

Thus, to rectify previous LD studies' inadequate attempts to consider and incorporate these four influential factors, the current research partially replicates the methodology used by Treffers-Daller et al. (2018), who illuminated LD measure

predictions of L2 general language proficiency under the condition of consistent text length based on the simple, lemma, and word-family counts. This dissertation investigates the operationality of written and spoken lexical diversity in predicting L2 proficiency in those related respective skills, such as IELTS-based writing and speaking. The dissertation seeks to validate different LD measures as L2 writing predictors while incorporating all the four factors. It will also examine LD measure predictions of L2 speaking proficiency when text length was duly adjusted based on the specific analysis units in order to find out the minimum text length required to achieve the efficacy of LD measures in predicting L2 speaking proficiency.

Chapter 2

Literature Review

2.1 Introduction

The literature on lexical diversity has long been validating the applicability of LD measures in second language (L2) contexts and has reported that we can indeed use LD measures as reliable L2 proficiency indicators. Several studies (Gonzalez, 2017; Jarvis, 2017; Treffers-Daller, 2013; Treffers-Daller et al., 2018; Yu, 2010) have identified multiple factors potentially influencing how LD measures can predict L2 proficiency to varying degrees, such as the word-counting criteria, L1 background, language proficiency, and text length. The following overview of the literature considers each of these different factors.

First, I consider different analysis units reflecting different inflectional and derivational knowledge levels, so LD measure predictions might be variable, depending on whether the analysis unit chosen suits individual L2 learners' word part knowledge.

Second, it seems inappropriate to assume that L2 learners from diverse L1 backgrounds might all have the same vocabulary range as Yu **(**2010) suggested in a study comparing learners of mixed L1 backgrounds with the learner groups of specific L1 (Philippines and Chinese) backgrounds.

Third, many LD studies investigate whether LD measures are discriminative with regard to different proficiency levels (e.g., CEFR B1 vs B2 level or beginner vs intermediate vs advanced levels). Still, only some have considered LD measure predictions of individual variation (i.e., between learners within a specific proficiency level).

Fourth, despite continuous efforts to overcome LD measure text sample size issues, even newly developed and robust LD measures, such as *D, MTLD,* and *HD-D,* remain influenced by text length. Only a few studies (Treffers-Daller, 2013; Treffers-Daller et al., 2018) have used a constant text length in assessing diverse vocabulary.

Further examination of the influences of these four important factors on LD measure predictions might contribute to current LD research and understanding. The twofold aim of the current chapter is to review the literature to date on this topic to examine the extent to which the LD validation studies reviewed have addressed such factors, and thus to create a foundation for the experimental (3, 4, 5, and 6) chapters that follow.

This literature review comprises three sections. The first section presents the research on the specific six LD measures, also to be used in the experimental chapters, from the most basic LD measures to the most recently developed robust measures. This section, in some detail, explains different LD measure calculations, as well as discusses their text length dependency, with a focus on two methods (ratio and text length standardization) in response to text length sensitivity.

The second section reviews ten LD studies investigating various LD measures in terms of their different validity (e.g., predictive, internal, convergent, or divergent validity) in the L2 context. This section summarizes the studies and provides commentary on their significance, as well as highlighting the weaknesses, with a special emphasis on LD measurement.

The third section is a brief review of four papers that have raised awareness of the criticality of the analysis unit choice in L2 vocabulary assessment. This section highlights why the most common unit, the word-family count, is only sometimes a good fit for L2 learners, and instead proposes alternative word-counting criteria that

might more accurately capture L2 learner vocabulary knowledge. In particular, the discussion is mainly based on the requirement to recognize the appropriateness of the flemma count, which lies somewhere between the lemma and word-family counts in Bauer and Nation's (1993) scheme.

The last section summarizes the extent to which each reviewed LD study attempts to address the above four factors and highlights the need to consider all these four factors together in validating LD measure predictions. The section briefly explains how these factors are to be controlled in the experimental chapters by partially replicating Treffers-Daller et al.'s (2018) paper, which points out the necessity of careful analysis unit choice in the LD research field.

2.2 Lexical diversity measurement and text length sensitivity

Lexical diversity, also known as lexical variation, is defined as "the range and the variety of vocabulary deployed in a text by either a speaker or a writer" (McCarthy & Jarvis 2007, p. 459). A wide variety of LD indices measure lexical diversity, ranging from the simplest LD measure, the number of different words (*NDW*), which is now termed *Types,* to more complex measures (e.g., *D*, *MTLD, HD-D*). LD measure calculations are based on the number of tokens (total words) and *Types* (different words) produced in the written or spoken samples.

Following Treffers-Daller et al. (2018), the experimental studies in chapters 3, 4, 5, and 6 compare the basic and sophisticated LD measures (the terms used by Treffers-Daller et al., 2018). The three basic measures comprise: (i) the most common and basic LD measurement (*Types*; the number of different words); (ii) an earlier attempt in response to text length concerns (*TTR*; Type-Token Ratio, Johnson, 1944) on which most LD measures are based; and (iii) a simple transformation of *TTR* (*Guiraud's Index,* Guiraud, 1954).

The studies also consider three sophisticated LD measures, claimed to be more resistant to text length variation: *D* (Malvern & Richards, 1997; Malvern et al., 2004), the Hypergeometric Distribution Diversity Index (*HD-D*; McCarthy & Jarvis, 2007), and the Measure of Textual Lexical Diversity (*MTLD*; McCarthy, 2005). I explain the calculations and text length sensitivity of these six measures in the following (2.2.1 and 2.2.2) sub-sections. These six LD measures to be used in the four experimental chapters are presented in Figure 2.1.

2.2.1 Basic LD measures

The most basic method of quantifying LD is a simple count of the number of different words (types) produced in spoken or written samples. For instance, the sentence, "*The white shirt is more expensive than the green shirt*", includes 10 tokens and 8 types since the words "*the*" and "*shirt*" appear twice. Words are repeated more in lengthier texts, so the number of different words (types) declines as text lengths increase. Since simple type counts are dependent on text sample size (e.g., Durán et al., 2004), LD scores of texts with varying lengths are incomparable (deBoer, 2014; Treffers-Daller, 2013).

A potential means of addressing the text length issue relates to the Type-Token Ratio (*TTR*), a calculation of the proportion of the number of types to the number of tokens in texts. *TTR* scores range from 0 to 1, with a higher value showing more diverse vocabulary. For example, two 50-word texts that include 25 and 30 different words receive *TTR* scores of 0.5 and 0.6, respectively, showing that a text with 30 different words is more lexically diverse. However, *TTR* remains sensitive to text length, with several papers showing that *TTR* scores decrease with longer texts (Koizumi & In'nami, 2012; Malvern et al., 2004; McCarthy & Jarvis, 2007; Zenker & Kyle, 2021).

An early attempt to reduce *TTR* text length sensitivity was *Root TTR* or *Guiraud's Index,* which calculates the proportion of the number of types to the square root of tokens (*Guiraud's Index*=types/√tokens). Although Zenker and Kyle's (2021) study exploring the minimum text length for different LD indices suggested that *TTR* simple transformation (*Guiraud's Index*) is not stable for any text lengths (50-200 words), Daller and Xue (2007) demonstrated the positive correlation of *Guiraud's Index* with text length.

2.2.2 Sophisticated LD measures

D (also known as *Vocd-D*), which can be calculated with the CLAN tool in the CHILDES system (MacWhinney, 2000) (https://childes.talkbank.org/), calculates the decreasing rate of *TTR* with text sample size. Gerasimos et al. (2015) explained that "estimating *D* involves a series of random text samplings to plot an empirical curve of *TTR* versus number of tokens for a sample" (p. 841). *TTR* values are calculated for 100 samples of 35 random words from the text, and a mean *TTR* value for all 100 samples is estimated. This process is repeated for 100 samples of increasing lengths (36, 37, 38 up to 50 words), resulting in 16 mean *TTR* values. The *D* values are calculated by presenting these values on a curve generated using the formula: $TTR=D/N * ((1 + 2 * N/D)^{1/2} – 1)$. *D* values can theoretically range from 0 to 120 with low *D* values demonstrating much word repetition and low lexical diversity in the texts.

More recently, two alternative measures have been proposed. First, McCarthy (2005) developed *MTLD*, a measure to capture how *TTR* decreases with token size. *MTLD* calculates "the mean length of sequential word strings in a text that maintains a given *TTR* value (here, .720)" (McCarthy & Jarvis, 2010, p. 384). They conduct text forward and backward calculations using this procedure, and the mean of the obtained

two values for these calculations makes up the *MTLD* value. Second, McCarthy and Jarvis (2007) proposed *HD-D,* a direct estimation of the probabilities of word occurrence in a text, in contrast to *D,* which is based on random text sampling and curve fitting. McCarthy and Jarvis (2010) then showed that *HD-D* can calculate the probability of any token for each lexical type in a random sample of 42 words drawn from a text.

McCarthy and Jarvis (2010) reported *MTLD*'s low correlations with text length and the flawed measure *TTR,* highlighting that *MTLD* is less sensitive to text length. Zenker and Kyle (2021) also reported that *D, HD-D*, and *MTLD* are stable across all varying text lengths (50-200 words). Although these sophisticated measures seem more resistant to text length variation than basic measures, text sample size to varying degrees has affected these measures. In response to this concern, researchers (e.g., Koizumi, 2012; Treffers-Daller, 2013; Treffers-Daller et al., 2018) suggest maintaining consistent text length when assessing LD. The current study's experimental chapters 3, 4, 5, and 6, therefore, analyze texts of the same length for reliable comparisons between LD scores.

Table 2.1

Basic and Sophisticated LD Measures and Calculations

LD Measures	Calculations
Types	Number of different words (V)
Type-Token Ratio (TTR)	Ratio of different words and total words (V/N)
Guiraud's Index (Guiraud)	Proportion of different words to the square root of total words ($V/\sqrt{N}$)
D	$TTR=D/N * ((1 + 2 * N/D)^{1/2} – 1)$
Measure of Textual Lexical Diversity (MTLD)	"Mean length of sequential word strings in a text that maintain a given *TTR* value (.720)" (McCarthy & Jarvis, 2010, p. 384)
Hypergeometric Distribution of Diversity (HD-D)	"Sum of lexical probabilities based on random samples of 42 words" (Nasseri & Thompson, 2021, p. 4)

2.3 Studies investigating LD measures' different validity aspects in the L2 contexts

LD studies have explored a variety of LD measures in relation to different aspects of validity, such as construct, internal, convergent, or divergent validities, and, in particular, their predictive validity. However, the studies differ in their participants, language modes, language tests, and LD measures. Gathering and assessing the existing LD studies are therefore necessary to acquire a thorough understanding of the current state of knowledge on LD measures' applicability in the L2 context. This section reviews ten significant LD studies by summarizing each study, making some judgments about their contributions to L2 vocabulary research, and identifying a few weaknesses that mainly relate to LD assessment.

This chronological order-based review section starts with Read and Nation's (2006) study, "Spoken lexical diversity and speaking proficiency", which examined the sophisticated LD measure, *D*, as a predictor of the IELTS speaking level along

with the lexical sophistication. Second, I review Yu's (2010) study, "Lexical diversity in writing and speaking task performances", which explored *Vocd-D*'s predictions of L2 general, writing, and speaking proficiency. Third, Jarvis's (2013b) study, "Defining and measuring lexical diversity" follows, with this study clarifying whether LD measures capture lexical diversity and have correlations with human's ratings of lexical diversity and writing. Fourth, McCarthy and Jarvis's (2013) study, "From intrinsic to extrinsic issues of lexical diversity assessment: An ecological validation study" is significant since the study examined LD measures' dependency on the textual and word count variations of a corpus. Fifth, Gonzalez's (2017) study, "The contribution of lexical diversity to college-level writing", follows with the study exploring LD measures' discrimination between groups and individual writing variations. Sixth, I review Jarvis's (2017) paper, "Grounding lexical diversity in human judgments", which is a heavily cited paper because of its attempt to address the insufficient construct validity of LD measures. Seventh, I review Treffers-Daller et al.'s (2018) paper, "Back to basics: How measures of lexical diversity can help discriminate between CEFR levels", investigating the extent to which LD measures predict general language proficiency, with a major emphasis on how the different analysis units can variously influence the LD measures. Eighth, Zhang and Daller's (2019) study, "Lexical richness of Chinese candidates in the grade Oral English examination", also validated LD measure discrimination between different speaking and general proficiency levels. Ninth, Nasseri & Thompson (2021), "Lexical density and diversity in dissertation abstracts: Revisiting English L1 vs. L2 text differences", validated the predictions of academic writing proficiency of the pairs or groups of LD measures with similar calculation methods. The last reviewed study is Zenker and Kyle (2021), "Investigating minimum text lengths for lexical diversity indices", which

explored different LD measures' text length dependency and the minimum length required for each LD index.

2.3.1 Read and Nation (2006): Spoken lexical diversity and speaking proficiency

Lexical diversity plays an important role in predicting both writing and speaking proficiency. Although written lexical diversity has been extensively examined (Jarvis, 2002; Treffers-Daller et al., 2018; Wang, 2014; Wu et al., 2019), spoken lexical diversity has received less attention. In responding to this comparative lack of research on spoken LD, Read and Nation (2006) analyzed the roles of lexical features, such as lexical variation (lexical diversity) and lexical sophistication in discriminating between the speaking proficiencies of IELTS test examinees. The authors qualitatively examined whether formulaic language use differed across various speaking levels.

Read and Nation analyzed a small spoken corpus of 88 examinees who had taken the actual 2002 IELTS Speaking Tests held at 21 test centers around the world, including in Australia, the United Kingdom, Columbia, New Zealand, Ireland, Peru, India, China, Hong Kong, Cambodia, Pakistan, Libya, and Sudan. Their proficiency ranged from levels 4 to 8. The authors examined eighty academic and eight general training texts for four different topics (*Eating out*, *Reading a book*, *Language learn*ing, *Describing a person).*

For data preparation processing, the speeches were transcribed by trained experts over the course of almost nine months, and the transcripts were treated so that the interviewers' utterances, pauses, and notes on speech quality were omitted through electronic editing. Then, from the resulting text files, the preparation processing was carried out manually: deleting hesitations, back-channeling utterances,

and false starts, separating the short forms, and rendering multi-word proper nouns as single words.

For their analysis, Read and Nation applied four quantitative automated measures (*Wordsmith, D, Range*, *P-Lex)* to calculate the lexical statistics and conducted exploratory and subjective analyses to qualitatively analyze the formulaic language use.

First, the authors conducted three analyses using the *Wordsmith* tool to examine the participants' overall lexical production, identifying the frequent content words related to four topics and keywords. The analysis showed the decreasing trend of type and token numbers from the advanced levels to lower levels, confirming that higher-proficiency learners had a larger and more diverse vocabulary than the less proficient learners. However, there were large variations within levels; for instance, within IELTS level 8, examinees' texts varied in length from 728 to 2741 words. The authors, therefore, concluded that the raw number of words could not predict proficiency levels. Second, topic variation analysis with the *WordSmith Wordlist* tool showed that "*Language learning*" generated the longest word list because of the speakers' similar vocabulary use based on their common language learning experiences. The topics "*Reading a book*" and "*Describing a person*" created the shortest wordlists because of the diverse choices of books or people. Third, the *WordSmith Keywords* analysis reported the association between the words that convey the salient meaning and each topic.

Second, Read and Nation employed an LD measure (*D)* to calculate the different words used in the speeches. LD score analysis showed that LD scores declined as the levels went down. High-proficiency speakers used more diverse vocabulary than lower-proficiency speakers. However, varying *D* scores at each

proficiency level indicated that *D* alone was not able to differentiate between the speaking proficiencies of the test-takers.

Third, the authors used *Range* to create the vocabulary frequency profiles. It classified the words deployed in the spoken texts into four lists: First 1000 most frequent words, Second 1000 most frequent words, Academic Word List, and Not in Lists. The analysis indicated that over 50% of the spoken words comprised the first 1000 most frequent words, with the percentage increasing with the lower levels. The proportion of the second 1000 most frequent words used at different levels seemed unstable. In terms of academic words, the spoken production of IELTS levels 6 to 8 comprised 9-10%, whereas level 4 texts represented the lowest proportion (6%). What was not in the three lists made up 21-12% from the highest to the lowest levels. Observing the Academic Word List and Not in Lists, the higher-level examinees used more sophisticated words than the lower-level examinees.

Fourth, Read and Nation used the *P-Lex* tool to examine whether there were differences in infrequent vocabulary use between proficient and less proficient speakers. The analysis showed that higher-level speakers received higher lambda values, implying examinees with higher speaking proficiencies used a greater number of sophisticated words.

Fifth, the authors carried out a qualitative analysis to identify the types of formulaic expressions used at three different speaking (4, 6, 8) levels. The findings illustrated that the formulaic language use across the levels was different. The most advanced speakers (level 8) used more word sequences, and low-frequency technical words and short words or phrases for pragmatic purposes. The level 6 speakers used a limited vocabulary range, fewer idioms, both appropriate and inappropriate word sequences or individual items, and no pragmatic devices; however, they could still

maintain meaningful and effective communication. The least proficient (level 4) speakers, though, mostly used high-frequency words and almost no formulaic expressions.

To conclude, Read and Nation reported that higher-proficiency speakers used greater vocabulary and more diverse and sophisticated words and more formulaic expressions than less proficient speakers. However, the text length, lexical diversity, and sophisticated measures alone could not predict the IELTS speaking bands.

Read and Nation's study is significant for two reasons: (i) useful insights into the vocabulary use across different speaking levels for the development of the IELTS lexical resource evaluation, which is one of the four speaking assessment criteria; and (ii) the exploration of the multi-word expression use of speakers at different IELTS speaking levels.

First, the authors reveal that speakers at varying proficiency levels use the vocabulary differently, including diverse and sophisticated words. There were also variations in vocabulary use within levels (e.g., different lengths within level 8 speakers, LD score dispersion within levels 6 and 7). These important findings raise awareness of the improvement of the descriptors to appropriately assess the vocabulary knowledge necessary for the specific levels and help the raters understand how they should evaluate the examinees' lexical knowledge.

Second, Read and Nation's paper is one of the few studies (Garner & Crossley, 2018; Kyle & Crossley, 2015; Tavakoli & Uchihara, 2020) that address multi-word units, which have been neglected compared to the study of single words used in the speaking samples. The finding of the different word sequence use between competent and less competent speakers highlights the important role of formulaic language use in predicting speaking abilities. Defining formulaic language, with the

careful consideration of the expression qualities, might support future formulaic expression studies.

Despite these two significant findings, the study has at least three limitations relating to lexical diversity evaluation: (i) varying text lengths; (ii) mixed L1 backgrounds; and (iii) lack of information on the word-counting technique.

First, the spoken transcripts used in the study differ in length. This might be a reason for the reported finding of the LD score differences not being significant between different speaking proficiency levels, despite the difference in diverse vocabulary use within and between levels. *D,* which seems to be less sensitive to the text sample size, has been improved as a response to the text length dependency of other LD measures, such as *TTR* and *Guiraud's Index*. However, the measure remains affected by text length, like other existing LD measures, so it would be more appropriate to keep text lengths constant in assessing lexical diversity.

Second, the examinees of the study came from various countries, and some candidates were even from English-speaking countries, such as Australia and the United Kingdom. Learners from different L1 backgrounds might have different lexical knowledge. The study's findings are based on participants of mixed L1 backgrounds, the findings seem less generalizable to specific learner groups (L1, ESL, or EFL).

Third, despite the careful treatment of the spoken transcripts, the information on the word counting criteria employed (e.g., the use of simple or word family count) is not mentioned. However, the analysis unit choice seems important, as learners might have different lexical (inflectional and derivational) knowledge. For instance, the use of a word family count might not discriminate between speakers with low

derivational knowledge. The study should have explained the analysis unit used both for clarification and to help navigate future research.

In conclusion, Read and Nation explored the different vocabulary use (e.g., lexical diversity, infrequent words, multi-word units) within and between speaking proficiency levels. However, for more reliable LD findings, the authors might have more carefully controlled text length and L1 background and focused on the method of counting different words.

2.3.2 Yu (2010): Lexical diversity in writing and speaking task performances

Researchers have examined either written or spoken lexical diversity to determine their accuracy in predicting general L2 language proficiency and specific writing and speaking abilities (Jarvis, 2002; Read & Nation, 2006; Treffers-Daller et al., 2018). However, what appears lacking is a single study exploring lexical diversity used in both the written and spoken production of the same participants as well as addressing LD measures' ability to discriminate between writing and speaking levels. To respond to this gap, Yu (2010) compared the diverse vocabulary used in both writing and speaking products of learners of L2 English and explored the extent to which LD measures predict overall language, writing, and speaking proficiencies. He further examined whether there are any lexical diversity differences in the spoken and written performances of the same participants.

The participants were 200 L2 learners who took the Michigan English Language Assessment Battery (MELAB) test between 2004 and 2005, mainly for college admission and professional certification. They were from 38 first language backgrounds, including four major groups (Filipino, Chinese, Russian, Persian). The archived data included 200 compositions on five different writing topics (personal and impersonal) and 25 spontaneous face-to-face interviews being tape recorded. The

compositions ranged from 123 to 735 words, and the interviews ranged from 210 to 2163 words.

In terms of data processing, Yu prepared both written and spoken data: handwritten texts were converted into MS Word and only careless spelling mistakes were corrected; spoken data were transcribed word by word, with the interviewers' utterances and non-words removed. For both written and spoken texts, Yu counted all the inflected forms of the words as different types as the initial word frequency analyses indicated the data included only a few inflections, and the MELAB test mainly requires learners' word form knowledge. Then *D* was computed with the *vocd* command in the CLAN program available at https://childes.talkbank.org/. Each written or spoken text was run 15 times, and the obtained scores were averaged to get the final *D* scores to maintain consistency since *vocd* generated slightly different LD scores, even for the same text, each time the program was run. As for the measures of other linguistic features, the *freq* command was employed to create a word list and find the word frequency, and the *WDLEN* and *WordSmith Tools* were used to compute the lengths of words and sentences by the numbers of letters used.

Yu conducted four analyses to investigate the relationships of lexical diversity to writing, speaking, and general language proficiency, and the correlations between written and spoken lexical diversity. First, regression analyses were performed to explore the extent to which LD and other lexical features (tokens, types, word and sentence lengths, and the number of long and short words) in compositions were predictive of writing proficiencies. The findings indicated D could predict 11% of the writing variances, so *Types* and the number of long words used seemed to be more effective than *D*. *Tokens* was equally effective as *D* in discriminating between writing scores, whereas word and sentence lengths and the number of short words used were

less powerful than *D*. Yu conducted an additional investigation on L1 influence on LD measures predictions. LD measures could predict the writing proficiency of the mixed L1 group but could not predict the writing variances for the two largest L1 (Filipinos and Chinese) groups.

Second, Yu investigated the relationship between spoken vocabulary range and overall speaking proficiency. Simple regression analysis revealed that the vocabulary range used in the interviews could strongly predict different speaking levels. Compared to other lexical measures, *D* was a better speaking predictor. Yu interpreted that the proportion of the variances that *D* predicted (23.4%) was high as there might be many other factors affecting humans' evaluation of speaking proficiency.

Third, the author explored whether written LD and spoken LD were correlated. The findings indicated almost no LD mean score difference (D_{comp}=76.29 and $D_{interviews}$=74.11), implying a similar degree of lexical diversity in the two separate (writing and speaking) skills of the same candidates. He argued *D* was more discriminative of speaking score differences (23.4%) than writing score differences (11%).

Fourth, he examined the correlations between the diverse vocabulary used in the written and spoken outputs and overall language proficiency. The findings showed that written and spoken LD were significantly correlated with the final MELAB scores that averaged the compositions, listening, and GCVR scores (grammar, cloze, vocabulary, reading). Yu therefore concluded that the participants' four skills (writing, speaking, listening, reading), grammar, vocabulary, and overall language proficiency positively correlated with each other.

In conclusion, Yu showed that written and spoken LD showed MELAB writing and speaking scores and general L2 language proficiency, confirming the predictive ability of the LD measure used in this study (*D*). He found strong and significant correlations between the written and spoken vocabulary ranges of the same participants, as well as between their different skills (four language skills, vocabulary, and grammar).

Yu's paper is significant for at least three reasons: (i) evidence of an LD measure, *D*'s usability in a different language (MELAB) test; (ii) examination of a single LD measure predictions of both writing and speaking; and (iii) awareness of possible L1 influences on LD measures predictions.

First, he shows that one LD measure, *D*, is a useful indicator of each of the general, speaking, and writing proficiencies of MELAB L2 test-takers. This finding supports the applicability of LD measures in various international language tests besides the most widely used tests, such as CEFR, TOEFL and IELTS. Second, his study appears to be a first attempt to examine an LD measure's ability to predict both the writing and speaking proficiencies of the same participants. The study provides the useful findings that the writing and speaking outputs of the same learners seem lexically diverse to a similar extent, and that the LD measure, *D*, is a stronger predictor of speaking ability than writing.

Third, Yu shows that L1 background can influence participants' different word use. Despite significant correlations between lexical diversity *D* and overall writing ability for the whole study, additional analyses indicated different results for the subgroups. *D* could not predict the writing scores of the Filipinos and Chinese although it could predict the writing scores for the other mixed L1 background learners. These findings highlight the possible L1 background influence on LD

measure predictions. With this important finding in mind, future research should carefully consider L1 background in LD assessment.

Despite these clear strengths in Yu's study, there are two potential issues relating to LD assessment which need further investigation. These relate to the varying text lengths again, and the examination of a single LD measure, *D*'s prediction of L2 language proficiency.

First, Yu examined compositions and interviews that vary in length. Since LD scores decline with increasing text lengths (McCarthy & Jarvis, 2007; Richards, 1987), it seems somewhat inappropriate to compare LD values of texts with different token numbers. As Treffers-Daller (2013) recommended, the author should have set a stable text length to get valid LD scores for more reliable comparisons. It would be better for future LD studies to consider keeping the texts constant to minimize LD measures' text length dependency.

Second, Yu provides the interesting finding that *D* indicates MELAB general, writing, and speaking proficiencies and is a more powerful discriminator of speaking ability than writing ability. However, the study might have yielded more informative findings if it could have examined and compared different LD measures and then extrapolated which specific LD measures are better for assessment of writing proficiency. Such comparisons of different LD measures may be helpful to guide the appropriate LD measure selection for assessing a specific language skill (e.g., writing or speaking). I therefore recommended more studies to validate different LD measures, not just a single LD measure.

To conclude, Yu's study reported *D* as a valid indicator of MELAB general language proficiency, and as a better predictor of speaking than writing; however, it highlights the need for further studies to analyze texts of consistent length to reduce

text-length variable impacts on LD measures' predictive accuracy and to compare different LD measures in assessing lexical diversity.

2.3.3 Jarvis (2013): Defining and measuring lexical diversity

LD measures have been validated in terms of their correlations with other constructs: lexical aspects, such as lexical sophistication (Read & Nation, 2006), as well as language proficiency such as writing (Treffers-Daller et al., 2018) and speaking (Zhang & Daller, 2019); however, what remains unclear is whether existing LD indices are truly measuring what they are supposed to measure (lexical diversity). Jarvis (2013) claimed that, at the time of writing, LD measures lack construct validity. The measures were not based on any theoretically grounded definition of what lexical diversity is, and the extent to which these measures actually assessed lexical diversity itself was unclear. The author, therefore, attempted to resolve such issues by defining lexical diversity, proposing a model that assesses its properties, and validating that model with human judgments of lexical diversity and writing.

To address the lack of LD construct validity, Jarvis (i) discussed whether lexical diversity was subjective or objective; (ii) identified the internal properties of lexical diversity and determined the effects of these properties on human perceptions of LD; (iii) proposed objective measures of each property; (iv) validated the measures with human ratings of lexical diversity and writing proficiency; and (v) examined whether human raters considered lexical diversity during writing assessment.

First, Jarvis addressed whether lexical diversity was an objective or subjective construct. He argued that existing LD measures, whose calculations depend on frequency counts of type and token, were precisely calculating the proportions of the novel words and repeated words used in a text but not truly measuring lexical diversity. He showed that two texts with different lengths (100 and 200 words) had

the same *TTR* scores (0.45) and of repetition (0.55) but different redundancy (use of words that are not required), implying that repetition could not explain redundancy. Unlike repetition, redundancy was perception-based, and it might be objectively measurable when the ways in which humans perceived redundancy were known. Similarly, the existing LD measures which were capturing lexical variability (as claimed by Jarvis) could not adequately measure lexical diversity based on human perceptions. However, lexical diversity, a subjective construct, could be assessed through theoretically strong objective measures that were developed and validated with human perceptions of lexical diversity.

Second, Jarvis identified what properties made up lexical diversity and the property effects on human perceptions of lexical diversity. He described six internal properties of lexical diversity: variability, volume, evenness, rarity, dispersion, and disparity. He investigated their influences on human LD judgments by using two tasks (paired-sentence and paragraph-sorting tasks). In the paired-sentence task, 130 participants (106 undergraduates, 18 graduates, and 6 others) rated the lexical diversity level of six sentence pairs, with each pair representing one property. Five properties except dispersion (average distance between each type number) were tested, but the disparity applied to two pair sentences rather than for formal and semantic disparity. The findings indicated human judgments affected four properties except disparity, and variability and volume were the most influential properties on human judgments. As for the paragraph sorting task, 38 undergraduate and graduate students from an American university classified a baseline paragraph and five paragraphs in which each property was applied from the highest to the lowest level of lexical diversity. The findings indicated that the students ranked the text with high rarity as the most lexically diverse text, followed by the longest text, and the text with

the largest type numbers. As expected, two texts changed with low evenness (token distribution across types), and disparity (lexical type differentiation degree) levels were less lexically diverse than texts modified with high property levels; however, the base-line text was judged as the text with the lowest LD level.

Third, Jarvis attempted to propose the objective measures that could capture each LD property. He suggested *MTLD*, which was less affected by volume and evenness than other existing LD measures, as the measure of variability; the total word numbers as the volume measure; the standard deviation of tokens per type as the simplest way to assess evenness; the mean lemma rank of the British National Corpus as the rarity (less frequent word use) measure; the mean distance between tokens of the same types as the dispersion measure; and the mean number of words with similar semantic sense as the disparity measure.

Fourth, the author examined whether the measures of LD properties were correlated with writing proficiency and with each other. The writing samples included 210 written texts of varying lengths (from 24 to 578 words) produced by L1 Finnish and L1 Swedish speakers of English who were in grades 5, 7, and 9. The texts were evaluated by two trained raters from Indiana University by using a 26-point rating scale. Because of the high inter-rater reliability, the rating mean scores were used. To calculate lexical diversity values, the essays were lemmatized by having the words converted into the related base forms, and then the types and tokens were computed with Perl software, the mean BNC rank was calculated regarding the BNC lemma file, and the word mean number per sense was calculated using the WordNet sense file. The findings indicated that five out of the six diversity measures could predict writing proficiency except rarity, and dispersion seemed the best predictor. The high

correlations of volume with evenness and dispersion suggested that evenness and dispersion measures depended on text length.

Fifth, the author performed a series of regression analyses to explore whether human raters considered the identified properties while evaluating lexical diversity. 50 texts of 7.5, 10, 13, 16, and 20 points on the 26-point rating scale were treated in which spelling and grammatical errors were corrected. The texts were assessed by eleven raters (eight ESL teachers and three graduate students) who had received no training besides the simple instruction to quickly read the essays and to assign scores based on 1-10 points. The findings showed that the LD measures seemed predictive of human LD judgments except for rarity, that volume was again highly correlated with evenness and dispersion, and that the six measures could capture fewer than 50% of the humans' LD rating differences. Aside from volume, a regression analysis of the five measures, showed that all measures could predict human LD ratings.

To conclude, Jarvis claimed LD to be a subjective construct comparing six inherent LD properties and suggested that most of these properties affected human LD perceptions and that the proposed measures are indicative of human judgments of writing and lexical diversity.

Jarvis' study is significant as it attempts to fill an important gap in LD the assessment field by defining LD, identifying its inherent properties, and proposing a potentially useful set of measures. His study raises awareness of the need for a clearer and more theoretical LD definition, on which the reliability and validity of the existing LD measures could be based and improved. His proposed six-dimensional model might be a good solution for the development of LD measures, based on a theoretical definition of LD.

Despite its significance, the study has at least two potential weaknesses: the need for an inductive study of human LD judgment, wider validation of different LD measures, and keeping text length constant.

First, the author claimed lexical diversity to be a subjective construct and examined whether human judges considered the identified intrinsic six properties of lexical diversity. However, such deductive validation might not capture the perceptions or intuitions of human raters to the full extent as it is limited to only six proposed properties. There might be other aspects that influence human judgments; future study should therefore inductively explore how, and what as human raters assess LD without reference to the presumptive LD properties.

Second, the author argued that the existing LD measures are not truly assessing lexical diversity. However, the LD measure used in his study, *MTLD*, has been found to be the property that most affects human judgments, and also it could predict overall L2 writing proficiency and human LD ratings. This reported finding of *MTLD*'s usability is in line with Gonzalez (2017) and Treffers-Daller et al. (2018). It is significant that we can still assume *MTLD* to be a useful measure in predicting language proficiency (e.g., writing) and the texts' lexical diversity. Researchers should explore how other existing LD measures are closely related to human lexical diversity judgments and language proficiency to prove Jarvis' argument.

In conclusion, Jarvis addressed the imprecise definition of the construct of LD, and proposed the measures validated with human LD intuition. However, it would be more informative and effective if future research could examine how human raters naturally incline to define and evaluate lexical diversity found in the spoken or written texts.

2.3.4 McCarthy and Jarvis (2013): From intrinsic to extrinsic issues of lexical diversity assessment: An ecological validation study

Researchers (Duran et al., 2004; McCarthy & Jarvis, 2010; Treffers-Daller, 2013; Zenker and Kyle, 2021) have attempted to address the intrinsic issues of lexical diversity measurement, such as the development and validation of more stable LD measures. However, the extrinsic issues, i.e., how LD measures work with a corpus that naturally varies in text lengths and word counts, remain unexplored. To respond to this shortcoming, McCarthy and Jarvis (2013) addressed LD measures' ecological validity (usability on naturalistic corpus data) by exploring the extent to which LD measures were reliant on the text count and word count variations of a corpus.

McCarthy and Jarvis analyzed a corpus of 276 English narrative texts written by L1 speakers of English, Finnish, and Swedish. The participants were classified into three groups of L1 English speakers (N = 66), four groups of L1 Finnish speakers (N = 140), and two groups of L1 Swedish speakers (N = 70) according to their grade levels and numbers of English language learning years. Written texts were holistically assessed by two trained raters from Indiana university by using the same 26-point scale used by Jarvis (2013), and the ratings were averaged. *MTLD, HD-D*, and *Maas* values were calculated using the Gramulator.

For data analysis, McCarthy and Jarvis performed the ANOVA, ANCOVA, and correlation tests to determine how the word count and text count variations of a corpus influenced LD measure predictions. First, the authors performed Pearson analyses to investigate the correlations between LD measures and their relationships with word counts. The first correlational analysis indicated that LD measures were highly correlated with each other. *MTLD* indicated strong correlations with *HD-D* (r=.830) and with *Maas* (r=.642), while *HD-D* and *Maas* were moderately correlated

(r=.586). The second analysis illustrated that word count had the highest correlation with *Maas* (r=.399), followed by *HD-D* (r=.369), and the lowest correlations with *MTLD* (r=.168).

Second, McCarthy and Jarvis performed ANOVA and ANCOVA analyses on different groupings of the participants**.** For the first grouping of three L1 English speaker groups, four L1 Finnish speaker groups and two Swedish speaker groups, an ANOVA analysis indicated that group membership significantly affected all three LD measures and word count. Following additional ANCOVA analysis, the authors reported significant correlations of word count with *HD-D* and *Maas*, but not with *MTLD*. Moreover, *MTLD* could discern most variances in LD scores across groups (22.7%) while *HD-D* and *Maas* could discern 20.0% and 19.7% of LD score differences.

As for the second grouping of L1 speakers of Finnish, Swedish, and English, the ANOVA analysis illustrated the significant main effects of group membership on word count, *HD-D*, and *Maas*, but not *MTLD*. Further analysis with the ANCOVA test showed significant correlations between word count and all three LD measures, with the effect size for *MTLD* being smaller than for *HD-D* and *Maas*. For the third grouping of L1 and L2 English speaker groups, ANOVA results indicated that grouping had little effect on LD measures and no effect on the word count. Additionally, this insignificant effect of third grouping was similar to an analysis of the second grouping of the three different L1 backgrounds, that grouping had high correlations with *HD-D* and *Maas* but low correlation with *MTLD*.

For three language groups (American, Swedish, and Finnish ninth graders), ANOVA analysis indicated that native-language grouping affected only *Maas*, suggesting that *Maas* scores varied across groups. Further ANCOVA analysis

confirmed that word count influenced only *Maas*. For three Finnish L1 speaker groups (fifth, seventh, and ninth graders), ANOVA results indicated that grade-grouping had significant influence on word count, *MTLD*, and *HD-D* but not *Maas.* *HD-D* could distinguish between all three grades whereas *MTLD* and word count could at least differentiate the lowest grade from the two higher grades.

In conclusion, McCarthy and Jarvis showed that *MTLD* and *HD-D* were less affected by the variations in word and text counts of a naturalistic corpus than *Maas*, and *MTLD* was the most reliable measure while *Maas* was the most sensitive measure.

McCarthy and Jarvis's study is significant as it fills an important research gap by determining LD measures' extrinsic issues (LD measures' usability on a naturalistic corpus with text length and word count variations), which had not been sufficiently examined. Naturally, written or spoken texts might differ in length because of writers' or speakers' differing language proficiencies, vocabularies, or content knowledge. It is also necessary to examine LD measures' applicability to such naturalistic data. McCarthy and Jarvis provided helpful information on the effective use of LD measures in corpus-based studies by suggesting the appropriateness of *MTLD* and of *Maas* depending on their degrees of sensitivity to corpus word count variations.

Despite such important information about the effects of corpus variation on LD measures, the study includes at least two major limitations concerning LD assessment: (i) no explanation of lexical unit analyzed; and (ii) the need for further analysis on other well-known LD measures, such as *TTR, Guiraud's Index*, and *D*.

First, like Read and Nation (2006), McCarthy and Jarvis's study did not provide information about the word counting technique; however, a simple count

might have been applied, which considers the different words used in the texts as different types. The reported findings might be different if the researchers had applied alternative word counting units, such as lemma and/or word-family counts. These different operationalizations of what makes up a "word" result in the application of different analysis units which might influence LD measures' predictive ability. To choose the suitable analysis unit for the particular participants under examination is important, and depends on their inflected and derived word knowledge. For instance, a word-family count might excessively estimate the overall L2 proficiency of learners with insufficient derivational knowledge (Brown, 2018; McLean, 2017). Instead, a lemma or flemma count, which demands only learners' inflectional knowledge, might be a more accurate predictor for such learners. More research was therefore needed to examine different word units' influences on LD measures for useful insights on the appropriate lexical unit selection for accurately evaluating vocabulary range.

Second, they limited the study findings to only newer LD measures (*Maas*, *HD-D*, *MTLD*). It would have been more informative if the authors had examined the different existing LD measures, including the most common basic measures (*Types, TTR, Guiraud's Index*) and the sophisticated measure, *D,* which have been extensively used in the LD research field. Future research should expand the study by comparing both basic and newer LD measures for more insights to advance the development of LD measures.

To conclude, McCarthy and Jarvis's study addressed the influences of text count and word count variation on LD measures for corpus-based LD analysis and proposed the use of *MTLD,* which shows the least sensitivity to these variations. The authors highlighted *Maas*'s high text length reliance. However, the study paid little attention to the importance of the analysis unit choice and leaves an important gap

(that of basic LD measures' usability for a corpus with differing text count and word count) that needs to be explored.

2.3.5 Gonzalez (2017): The contribution of lexical diversity to college-level writing

There are indications that both lexical frequency (a word's frequency level) and lexical diversity (different word use in a spoken or written text) are correlated with the writing proficiency of high-proficiency L2 learners of multiple L1 backgrounds (Crossley & McNamara, 2009; Crossley et al., 2011; Douglas, 2016; Treffers-Daller et at., 2018). Despite such rich evidence of lexical diversity (LD) measure predictions of writing proficiency, lexical frequency (LF) measure predictions of writing proficiency remain unclear. Gonzalez (2017) therefore attempted to investigate the relationships between lexical diversity and lexical frequency measures and evaluate which measure is a better L2 writing proficiency indicator.

The author analyzed two different learner groups. One was a monolingual English-speaking (MES) writer group comprising 68 students enrolled in a first-year composition course at a US public university. The second was a multilingual (ML) writer group comprising 104 L2 learners of English who were enrolled in the advanced second language writing courses of US-based Intensive English Programs (IEPs). They were from 13 L1 backgrounds, including Japanese, Thai, and Ukrainian, with the majority being L1 Spanish (N = 45), and Arabic (N = 36).

For data collection, MES writing samples of varying lengths (501-2530 words) were taken from four sections of the first-year writing course. ML writing samples (102–1003 words) were collected by using various US-based English L2 writing and English teaching listservs, electronic mailing lists where the topic creation and discussions between the subscribers were available. The samples covered the

seven different academic writing genres of analysis, cause and effect, compare and contrast, narrative, persuasive, process, and summary. Written texts were evaluated by three trained raters, using the TOEFL iBT independent writing rubric, and were then assigned TOEFL bands ranging from 2 to 5.

Data processing included the correction of the minor errors (spelling, repeated consecutive words, extra spaces, or punctuation marks). Gonzalez employed two indices to compute lexical scores: (i) an LD measure, *MTLD*, and (ii) an LF measure, *CELEX* (Baayen et al., 1995), which is based on a lemma count. The author used MANOVA, Pearson product moment, and binary logistic regression analyses to examine these LD and LF measures as the writing predictors.

First, MANOVA analyses were conducted to find the differences between the MES and ML writer groups in terms of lexical diversity and frequency. The findings showed that the written texts produced by L2 learners included fewer diverse words and more high-frequency words compared to the MES texts. Both measures could discriminate between these two learner groups; however, *MTLD* ($\eta 2 = .25$) was a stronger discriminator than *CELEX* ($\eta 2 = .25$). Regarding intragroup discrimination, *MTLD* could discern within-group variations in both MES and ML writer groups, whereas *CELEX* was effective only for ML writer intra-group variations. Second, Gonzalez performed Pearson product moment analysis to explore the correlations between lexical diversity and frequency in compositions. The findings indicated the moderate use of lower frequency words in essays with a wider vocabulary range ($r = -.44$).

Third, the examination of LD and LF measures predictions of writing proficiency using the Binary logistic regression analyses showed that only *MTLD* could significantly contribute to the regression model for both ESL and ML writer

groups. Based on MANOVA analyses, *MTLD* was found more indicative of writing score variances. The lexical range used in the essays of TOEFL band scores 4 and 5 was higher than in the essays at the two lower levels (TOEFL scores 2 and 3), which indicated differences in lexical frequency but not in lexical diversity.

To conclude, Gonzalez's study illustrated that the advanced multilingual L2 learners used more higher frequency words, and fewer different words in their written texts in comparison with the monolingual L1 English speakers. The study supported the previous findings of LD measures' usability in predicting overall L2 writing proficiency.

Gonzalez's study is important for two main reasons: (i) validation of an LD measure predictions of writing proficiency in comparison with a lexical frequency measure; and (ii) the exploration of both inter- and intra-group lexical diversity and lexical frequency variations for specific learner groups (ML and MES groups).

First, the author compared the two important vocabulary aspects of lexical diversity and lexical frequency and examined whether lexical diversity plays a greater role in predicting writing quality compared to lexical frequency. The finding that the LD measure was a more useful writing indicator supports the previous findings of LD measure predictions. As the author stated, this finding suggests the need for guiding learners to use different words in their written production, since the use of vocabulary range has a greater impact on writing scores and confirms LD measures' predictive validity.

Second, most studies have been concerned with exploring the extent to which LD measures predict between-group lexical or proficiency variations. However, Gonzalez added deeper insights by exploring LD measures' capability to predict both between- and within-group lexical and writing score differences.

Despite making these two valuable contributions, the study incudes at least three weaknesses that could impact the generalizability of the LD assessment findings: the word counting method, the first language backgrounds of the participants, and the varying text length. The first limitation relates to the way of counting different words. The study used *CELEX*, which is lemma-based, in measuring lexical frequency but cannot explain how a "word" is conceptualized in this LD assessment. This again seems to imply the use of a simple count, which considers all words that appear once in texts without repetition as different types. The study should have focused more on how a "word" is operationalized to enable readers to clearly comprehend the analysis unit.

The second limitation is that participants were from a wide variety of first language backgrounds. Gonzalez's findings of the significant intragroup LD score variations for the ML group appear to support Yu's (2010) claim of the potential impacts of L1 background on LD scores and measures. The productive vocabulary (vocabulary range) use of the L2 learners in the study, who were of various nationalities, might be different. The study should have determined whether we can find the same results with each specific L1 background.

Third, the study did not consider the importance of consistent text length in assessing LD. The writing samples vary in lengths (102 –1003 words for ML writers and 501-2530 words for MES writers). Because of the prominent and obstinate issue of the impact of varying text length on LD measures, the different word numbers are likely to decrease with longer texts; therefore, it is advisable to control for text length. For instance, the number of unique words used in the shortest length (102 words) text and the number of diverse words used in the longest length (1003 words) text seem

incomparable. For reliable LD score calculation and then comparisons, the study should have analyzed samples with the same token numbers.

To conclude, Gonzalez showed that the lexical diversity used in the writing samples was more discriminating than lexical frequency in predicting L2 writing proficiency, and that the LD measure, *MTLD,* could discern both between- and within-group writing differences. However, if the study could have addressed the aforementioned three factors which have been shown to have a pronounced impact on LD scores and measures, the findings of LD measure predictions might be more reliable.

2.3.6 Jarvis (2017): Grounding lexical diversity in human judgments

Despite the widespread use of LD measures in language assessment, there has been a concern about the lack of LD measure construct validity, as highlighted in the literature review's evaluation of Jarvis (2013). Jarvis (2017) has continued to argue that existing LD measures have not only measurement-related problems (e.g., word counting unit, text length) but also a construct-related problem (inadequate LD construct definition). To address the problem of the vague or inadequate construct definition, Jarvis explored whether human LD judgments were sufficiently reliable to be of a standard to which LD measures could be evaluated and validated against.

Jarvis analyzed a corpus of 276 narrative English texts written by 140 L1 Finnish speakers, 70 L1 Swedish speakers and 66 L1 English speakers, all elicited in response to a short silent Charlie Chaplin film. Two experienced judges rated the written texts using the same placement test rating scale as for the intensive English program (IEP) at Indiana University, and the final scores were assigned on the same 26-point scale from Jarvis' previous studies. From the entire collection texts, 50 texts of different lengths (24 to 578 words) were selected from the rating levels at which

the participants' texts were most evaluated: Rating 13 (N = 20), Rating 16 (N = 10), Rating 20 (N = 8), Rating 10 (N = 7), and Rating 7.5 (N = 5), which comprised a fair mixture of all three L1 speaker groups. Prior to analysis, the selected texts were treated so that spelling and major grammatical errors were corrected, and function words (articles and prepositions) were added where necessary.

As for the analysis, Jarvis first explored the LD judgment inter-rater reliability between the 2011, 2012, 2014, and 2015 ratings by separate groups of English-proficient raters from Ohio University. Second, he conducted a correlational analysis to investigate the relationships between the 2014 and 2015 human LD ratings and three automated LD measures (*HD-D, MTLD, MATTR*). Based on the analysis findings, Jarvis developed a corpus-specific automated LD measure.

First, in 2011, eleven raters (eight instructors, including two ESL teachers and three graduates) rated the diverse vocabulary used in 50 written texts: 30 of the texts were assessed by three raters, and two raters evaluated 20 texts. Pearson correlation analysis indicated that the inter-rater reliability levels ranged from (r = -.02 to .65), and the mean value was low (r = .30), suggesting that human raters could not provide reliable LD ratings.

Second, in 2012, Jarvis used a greater number of raters (20 students) from Ohio University, (four upper-level undergraduates and sixteen MA students). The raters were given instructions (e.g., to read the text quickly, and to focus on lexical diversity, not on writing quality) for giving consistent ratings as well as a sample text at a LD rating level 5. Each rater assessed the 50 texts. The analysis showed that the correlations between the raters varied from (r = -.01) to (r = .76), and again the mean inter-rater reliability (r = .32) was low. However, in terms of overall reliability, Cronbach's value (α = .90) suggested that the raters had assessed LD in similar ways.

Third, in 2014, 21 raters (seven upper-level undergraduates and fourteen MA students) assessed the writing quality of the 60 texts using the CEFR-rating scale and one week later, they assessed the lexical diversity of the texts. This time, Jarvis encouraged the raters by awarding some points for the rating task and consistency. The author examined the correlations between the writing judgements as well as between the LD judgements and further explored whether writing and LD ratings were correlated. First, the writing score analysis illustrated that the correlations ranged from (r = .15) to (r = .81), and both the mean value (r = .54) and Cronbach's alpha value (α = .96) were high, confirming the similarities between writing judgments. Second, for the LD rating analysis, the correlation coefficients were from (r = .21) to (r =.84), and the mean r (.57) and the Cronbach's value (α = .96) proved the high LD rating consistency between raters. Third, the Pearson analysis with writing and LD ratings illustrated the high correlation between them (r=.89).

Fourth, in 2015, 20 students (eight undergraduates, one PhD student, and eleven MA students) rated both the writing and vocabulary ranges used in the 60 texts. The raters were provided with the same instruction and incentives given in 2014. The correlational analyses indicated that the mean LD ratings (r = .51) and the Cronbach's value (α = .96) confirmed the 2014 results of the high inter-rater reliability in LD ratings and suggested that the raters had indeed separately assessed LD irrespective of writing quality. The raters were found to have made consistent writing judgments (r = .49 and α = .95), and the analysis also suggested that LD and writing evaluations were highly correlated (r = 89).

Fifth, Jarvis explored whether the 2014 and 2015 LD ratings were correlated with three automated LD measures (*HD-D, MTLD, MATTR*), and the findings showed low to moderate correlations between human LD judgments and LD measures.

Based on the findings, Jarvis suggested ways to improve automated LD measures so that they would have greater construct validity. The proposed method included rating the lexical diversity of a corpus subset by enough human judges (e.g., 20 raters), identifying the factors affecting human LD perceptions to define LD theoretically, creating objective measures reflecting LD properties, evaluating the developed measures with human LD judgments, putting the indices into a model to weigh the factors, and applying the developed corpus-specific measures to the whole data.

To conclude, Jarvis showed different consistency levels in LD ratings (low consistency in the 2011 ratings, slightly higher consistency in the 2012 ratings, and high consistency in the 2014 and 2015 ratings). The author highlighted the high correlations between 2014 and 2015 LD ratings and suggested a model to measure the LD of the corpus data.

Jarvis' study is significant as it attempts to fill in the missing information on LD measures' construct validity by examining human LD perceptions and proposing effective LD measures using the insights gained for corpus use. First, the study showed that human raters can assess lexical diversity consistently even though there is no training or rubric given. This interesting finding suggests that human LD judgments might be a reliable standard against which to validate LD measures' accuracy. Second, Jarvis's proposed method of developing LD indices based on a theoretically sound LD definition might solve consider the LD's lack of construct validity.

Although the study provides this useful information for the development of construct-based LD measures, there remain at least two key issues in relation to LD assessment. As with other studies examined in this literature review, the study needs

to consider the importance of the analysis unit selection and LD measures' dependence on text length. It also leaves the gap of not having examined other LD measures beyond *HD-D, MTLD*, and *MATTR*.

First, Jarvis did not clearly explain how he defines "word". As mentioned above, different word definitions can cause the application of various word counting units that can then differently reflect learners' different word part (inflections or derivations) knowledge. Because of the emerging research evidence of the cruciality of the analysis unit choice in LD assessment (Treffers-Daller, 2013), it would have been of greater value to have focused on one specific unit or several specified analysis units to make more of a contribution to the current LD knowledge and to guide future research.

Second, the author analyzed texts of varying lengths. Despite the various attempts to overcome the text length issue, LD measures remain sensitive somehow to text length. Setting a constant text length is suggested as the safest way to calculate and compare LD scores. We should not ignore the fact that text sample size is a key factor affecting LD scores and LD measures predictions of L2 language proficiency.

Third, Jarvis argued that LD measures lack construct validity, possibly based on the reported results of weak correlations between human LD ratings and the three automated measures (*HD-D, MTLD, MATTR*) under study. However, the findings on the relationships between human LD judgments and other LD measures (e.g., *Types, TTR*) might be different. Therefore, future studies should consider validating a variety of automated LD measures, including both traditional and newer measures, with human LD ratings for stronger evidence.

In conclusion, Jarvis indicated that human raters have similar intuitions about written lexical diversity, and suggested how automated LD measures should be

developed by calibrating them with the reliable human LD ratings. However, further studies should aim to confirm his argument on the lack of construct validity of the other existing LD measures and should conduct more systematic lexical diversity evaluations with the focus on the analysis unit impacts on LD measures and the fixed text lengths.

2.3.7 Treffers-Daller et al. (2018): Back to basics: How measures of lexical diversity can help discriminate between CEFR levels

Existing LD measures have been sufficiently validated by both empirical and corpus-based LD studies relating to different validity aspects (e.g., predictive, internal), and possible L1 background effects, as well as the applicability to different languages (e.g., French, English, Cantonese). However, a recent concern has arisen about the analysis unit choice influence on LD measures' predictive power. Despite Treffers-Daller's (2013) findings that the analysis unit choice (a lemma count) has an influence on LD measures predictions of L2 language proficiency, the influences of different word counting units on LD measures remain unexplored. To reveal this important missing information, Treffers-Daller at al. (2018) conducted a broader study with simple, lemma, and word-family counts by exploring how these three units affected LD measures in predicting general L2 language proficiency.

The participants were 179 L2 learners of English from 47 different L1 backgrounds, and their general language proficiencies were from B1 to C2 levels at CEFR (Common European Framework of Reference). They wrote timed essays on one of two different topics as part of the Pearson Test of Academic English, which provided their total scores and vocabulary and writing scores.

Prior to beginning analysis, Treffers-Daller and her colleagues conducted data cleaning, lemmatization, setting of constant text length, and LD score calculations.

First, proper names, acronyms, cardinal numbers, and non-existent words were excluded, and spelling mistakes were corrected. Second, they created three text versions using three lemmatization principles: (i) no lemmatization (word type; different words as different types); (ii) lemmatization 1 (lemma; inflections and derivations as separate types); and (iii) lemmatization 2 (word family; inflections and derivations as the same types). Third, the authors set the stable text length (200 words from the middle of the essays) using the Gramulator. Fourth, the authors adopted six different LD measures: three basic (*Types, TTR, Guiraud's Index*) measures and three sophisticated (*D, HD-D, MTLD*) measures, and computed the written LD scores using the Gramulator, SPSS and CLAN programs.

Treffers-Daller et al. investigated the relationship between written lexical diversity and general language proficiency and explored the degree to which the three different lemmatization techniques (no lemmatization, first lemmatization (lemma), and second lemmatization (word family) influence LD scores and LD measures' predictivity of overall CEFR language scores.

First, the study showed that lemmatization affected LD scores and LD measures' predictive power. LD scores were the highest when no lemmatization was applied, followed by the LD scores from the lemma-based calculation, and then the LD scores obtained through the word-family-based calculation. Moreover, the LD scores were consistent across different levels, indicating that higher-level learners had greater lexical diversity. Among the three lemmatization principles, with the higher effect sizes, counting the inflected and derived forms of the words as separate types (i.e., a lemma count) could best enhance most LD measures' ability to discriminate different overall CEFR language levels. On the other hand, *HD-D* appeared the strongest discriminator on the raw data where uses of every different word were

different types, whereas *MTLD* seemed the most powerful based on the word-family count. These findings suggested that LD measures were more powerful discriminators of general language proficiency when based on a lemma count rather than word-family counts, which were challenging for most L2 learners because of their insufficient derivational knowledge (Brown, 2018; McLean, 2017).

Second, for repeated measures they conducted ANOVA analysis to determine LD measures' discrimination between CEFR levels based on the lemma count, with which most LD measures could increase their discriminating power. All three basic measures (*Types, TTR, Guiraud's Index*) could differentiate the lowest level (B1) from the three higher levels (B2, C1, and C2). Among the sophisticated measures, *MTLD* could predict the lowest level (B1) and the two highest levels (C1 and C2) while *D* and *HD-D* could discriminate between the lowest and the highest levels (B1 and C2). Of all LD measures, *TTR* with the highest *F*-value (18.923), was the strongest indicator of L2 language proficiency, and the greater *F*-values implied that all three basic LD measures appeared more powerful in discriminating between overall CEFR scores than the sophisticated LD measures.

Third, correlation and regression analyses were performed to investigate LD measures' correlations between themselves and with the Pearson scores (overall, vocabulary, and writing scores). First, the correlational analysis indicated strong and significant relationships between LD measures, particularly, basic LD measures, as well as the positive correlations of the LD scores with the Pearson scores. Second, regression analysis on the Pearson scores and the two most correlated LD measures (*Types* and *MTLD*) illustrated that *Types* was the better predictor of the Pearson scores and overall scores. This suggested that the simplest LD measure (*Types*) seemed more

useful than both the other mathematically formulated measures and the newer complex measures in their study.

In conclusion, Treffers-Daller et al.'s study highlighted that the lemma counting technique had a greater influence on LD measures compared to simple type and word-family counts, and it claimed that LD measures were reliable indicators of general CEFR language proficiency.

Treffers-Daller et al.'s study has made a large contribution to lexical diversity research for three significant reasons: (i) more systematic LD assessment through careful data cleaning and the use of controlled text lengths; (ii) exploration of the effects of different word-counting methods; and (iii) comparison of multiple LD measures.

First, the authors investigated lexical diversity in which the data were cleaned by removing the words that do not represent learner vocabulary knowledge (e.g., proper names, cardinal numbers, non-existent words). They weigh the effects of LD measures' dependence on text length, and thus analyzed texts of the same length while ensuring the inclusion of the beginning, middle, and concluding parts of the essays. These two processes might bring greater reliability on the study's findings as it responded to some research limitations of the previous LD studies.

Second, their study seems to have been the first attempt to test different word counting techniques in LD assessment and to explore their influences on LD measure predictions. Their analysis of three common word-counting units (simple type, lemma, and word-family counts) showed that lemma-based counting could best manifest most LD measures' predictive power, suggesting that the lemma count is a more discriminating lexical diversity measurement than either simple type count or word-family count for the L2 English learners in their study. The study highlights

how important defining precisely what is counted as “a different word” is in LD assessment and how LD measures’ predictive power varies depending on the analysis unit selected for use.

Third, the authors compared the predictive power of a variety of LD measures (basic and sophisticated) in a single study. The findings provided clear information about the extent to which various LD measure predictive abilities are different and which measures are the better measurements of general L2 language ability at CEFR levels.

Despite these three significant points, there remain three important issues that might need further investigation: (i) the relationships between written lexical diversity and writing proficiency; (ii) the examination of an alternative word count unit (the flemma, i.e., the inflections under different word classes as the same types), and (iii) the exploration of the effects of L1 background on LD scores and measures.

First, Treffers-Daller et al. examined the ability of the lexical diversity used in the compositions in discriminating between overall language scores based on combining all four language skills (listening, reading, writing, speaking); however, they failed to investigate the written lexical diversity’s relationship with writing quality alone. It is unclear how the other language skills relate to and influence writing; therefore, future studies should investigate whether the diverse vocabulary used in the written text can be an accurate predictor of the L2 writing quality.

Second, the authors made a wider investigation of the lexical units in LD assessment by comparing three common word-counting units (simple type, lemma, and word-family counts). Generally, L2 learners seem to have sufficient inflected knowledge but only have limited derived knowledge (Brown, 2018; McLean, 2017). Therefore, a lemma count, which considers the inflections under the same word class

might underestimate the participants' actual inflectional knowledge, whereas a word family count that entails a wide variety of affixes might overestimate their derivational knowledge. For this reason, the study should have analyzed a potentially more useful word counting unit (the flemma) for L2 learners as an alternative to the lemma and word-family counts. A flemma count assumes that the learners know the inflections regardless of the parts of speech, and thus represents a slightly higher level (level 2.5) than a lemma count (level 2) on Bauer and Nation's (1993) scale. It would be of great value to examine whether a flemma count is a better unit for assessing L2 English diverse vocabulary, potentially resulting in the stronger predictive powers of LD measures, compared to other units.

To conclude, Treffers-Daller and her colleagues attempted to evaluate lexical diversity more systematically and highlighted the variability in LD measure predictions of L2 language proficiency according to the different analysis units used. However, the study does not examine the correlations between this written lexical diversity and its related skill (writing), and it leaves the important gap of neglecting to examine the potential usability of the flemma count in assessing L2 lexical diversity.

2.3.8 Zhang and Daller (2019): Lexical richness of Chinese candidates in the graded oral English examination

Several studies (deBoer, 2014; Ha, 2019; Kyle and Crossley, 2015; Lai and Schwanenflugel**,** 2016; Lu, 2012; Treffers-Daller, 2013; Treffers-Daller et al., 2018) have validated existing lexical measures under different conditions (e.g., different learner groups, tests, or language modes). Despite the sufficient research evidence of LD measures predictions of writing proficiency, we need more research to explore LD measure predictions of speaking proficiency.

Therefore, Zhang and Daller (2019) examined the lexical richness (lexical diversity and sophistication) of Chinese candidates at different levels of a graded oral examination. Furthermore, the authors explored the relationship between lexical richness and speaking proficiency and with their general L2 proficiency. Finally, the authors interviewed the examiners for a deeper understanding of the candidates' performance and interactions during the oral tests.

Zhang and Daller analyzed the spoken data obtained from the 2008 Graded Examination in Spoken English (GESE) sponsored by Trinity College, London, and administered by trained Chinese local examiners. The study examined 158 candidates at three different grades in GESE (Grade 2 in the initial stage, Grade 5 in the elementary stage, and Grade 7 in the intermediate stage), and their language proficiency levels were equivalent to CEFR A1, A2-B1, and B2 levels. The candidates completed different interview tasks based on their levels. The Grade 2 candidates did a conversation task in which they performed some actions by following the instructions, the Grade 5 candidates talked about a topic and responded to the examiner's questions, and the Grade 7 candidates presented on a topic in an interactive style.

Prior to the data analysis, the audio recordings were randomly selected, transcribed by the CHAT format of the CHILDES Language Data Exchange System, and evaluated by 23 experienced local examiners. The spoken texts varied in length (66 to 482 words). The authors computed the lexical richness (*Tokens, Types, D, Guiraud,* and *Advanced Guiraud*) scores and calculated the Mean Length of Utterances (*MLU*) to measure the candidates' general language proficiency.

Regarding the lexical richness scores of the candidates at different grades, the analysis indicated that the *Tokens, Types, D,* and *MLU* scores tended to increase with

the higher grades. There were significant differences in all scores between the lowest grade (2) and the two higher (5 and 7) grades; however, only the *Advanced Guiraud* and *MLU* scores were significantly different between grades 5 and 7. All measures except *Advanced Guiraud* could discriminate between the Grade 2 candidates and the Grade 5 and 7 candidates who passed the test, whereas only *Advanced Guiraud* could distinguish between the Grade 5 and 7 candidates. Regarding the LD measure predictions of oral proficiency (Pass and Fail), all measures could discriminate between the Pass and Fail groups at Grade 2; however, *MLU* could not distinguish the Pass and Fail Grade 5 candidates, and *Guiraud's Index* and *MLU* were not discriminating of the Grade 7 candidates' Pass and Fail status.

For the qualitative data, three senior grade 7 examiners were interviewed on the candidates' performances and interactions. The examiners observed the sometimes-poor communication performance, in which some candidates frequently made long pauses and irrelevant responses, resulting in ineffective communication or even communication breakdowns. The examiners therefore argued that most Grade 7 candidates' actual proficiency levels did not fully match their chosen GESE grade. The examiners mentioned that a reason might be the prestige and prejudice associated with that grade classification. The candidates could choose any GESE grade, so most candidates enrolled in Grade 7, despite their low proficiency, because the Grade 7 certificate was important for enrollment in a prestigious secondary school. The authors also suggested another reason that the task type, which required more preparation time and greater lexical diversity knowledge, might have influenced the candidates' oral performance.

To conclude, Zang and Daller's study indicated that the elementary and intermediate GESE level candidates had higher lexical scores than the initial level

candidates. However, most Grade 7 candidates' lexical scores were not as high as the Grade 5 candidates' and did not meet the required proficiency for effective communication. The study suggested that all lexical measures except *MLU* were valid discriminators of GESE speaking proficiency of the Pass and the Fail groups at initial (Grade 2) and elementary (Grade 5) levels while *G* and *MLU* were not discriminative of the (Pass and Fail) candidates at the intermediate level (Grade 7).

Zang and Daller's study is significant for three major reasons: (i) providing evidence for lexical richness measures, including LD measures, and predictions of speaking proficiency; (ii) using random sampling from a large population; and (iii) raising awareness of the weak grade classification of the GESE test.

First, the study provided useful information on how measures of lexical diversity in speaking are applicable as GESE speaking proficiency discriminators (Pass and Fail). *Types, Tokens, D,* and *Advanced Guiraud* could predict the outcomes of all three groups at any stages, although *MLU* was not discriminative of the Grade 5 and 7 candidates' groups. The findings support the existing evidence of lexical diversity measures as valid speaking predictors, highlighting the greater importance of recognizing diverse word use rather than advanced word use in estimating the speaking proficiency of low-proficiency Chinese learners.

Second, the study randomly selected candidates from a large data set (Grade 2, 5, and 7 candidates from the 2008 GESE oral test). The findings of this random sample based research are more likely to be generalizable than other specific classroom based research.

Third, the study highlighted the weak classification system of the GESE test by indicating the Grade 7 candidates' lower oral proficiency and vocabulary

knowledge, meaning that they appeared to be less qualified than necessary, resulting in a low pass rate.

Despite these significant findings, the study had at least three potential limitations which should be carefully addressed in future vocabulary studies for more valid findings. These include the need for better data cleaning prior to the LD score calculation and analysis, and for close attention to both the text lengths and the word counting units used.

First, the study does not mention how the data in the spoken transcripts were cleaned during the data collection process. In lexical diversity assessment, data cleaning (i.e., removal of proper names, cardinal numbers, repeats, back channelling phrases) is important for producing reliable LD scores. Treffers-Daller (2013) and Treffers-Daller et al. (2018) have suggested deleting the words that might cause the inflation of LD scores. Therefore, future studies should carefully treat the data prior to analysis.

Second, the authors analyzed spoken samples of different numbers of words (66 to 482 tokens). Treffers-Daller (2013) argued that comparing the lexical diversity scores of texts of different lengths seems misguided. Supposing that both 50-word and 70-word samples included 25 types, it would seem unfair to compare these two texts. Moreover, the candidates in the study are low-proficiency L2 learners, so the longer speeches might include more repeated words. It will therefore be necessary to set a constant text length, depending on the LD measures used, as suggested by Zenker and Kyle (2021), who examined the minimum text lengths required for different LD measures to be effective.

Third, the authors did not explain how they count different words, so it is unclear whether they used a simple count (different words as different types), a

lemma count (headwords and related inflections under the same word class as the same types), or a word-family count (headwords and both inflections and derivations as the same types). There is increasing evidence of the importance of the word unit selection in LD assessment to enable the greater L2 language proficiency predictions using LD measures. For instance, the use of a simple count might underestimate high-proficiency learners' lexical knowledge, whereas a word-family count might overestimate low-proficiency learners' derivational knowledge. It is therefore necessary to choose the word counting unit that best matches learners' existing word part knowledge or proficiency, especially in the assessment of lexical diversity.

To conclude, the study confirmed the LD measures' predictive validity by showing that LD measures can be used as the GESE speaking proficiency discriminators of Chinese candidates of various levels. However, the findings would be more contributive to the LD field if the study had focused on data cleaning, LD measures' sensitivity to text length, and the influence of the analysis unit used.

2.3.9 Nasseri and Thompson (2021): Lexical density and diversity in dissertation abstracts: Revisiting English L1 vs. L2 text differences

Several studies have validated the use of the existing LD measures in L2 language assessment, with most attention focused on a single LD measure validation and/or comparisons between different LD measure predictions. Only few studies to date have examined whether LD measures produced better predictions when deployed as combined measures rather than using a single measure. One of these limited studies is Treffers-Daller et al. (2018), which tested the combination of one basic and one sophisticated LD measure in assessing general language proficiency. However, Nasseri and Thompson (2021) conducted a wider and more systematic study by examining lexical density and lexical diversity in academic written texts. The authors

investigated the learner group (L1 and L2 English postgraduate students) to access and compare the ability of fourteen LD measures, deployed in pairs and groups based on their similar calculation methods, to discriminate between the learner groups.

Data comprised 210 Master's dissertation abstracts written by three different learner (EFL, ESL, and L1 English writers) groups. The EFL writers (N = 70) were Iranian university students from different ethnic and L1 backgrounds, the ESL writers (N = 70) were Iranians studying in the UK for their Master's degrees, and the L1 English writers (N = 70) were British university students. The abstract lengths ranged from 175 to 300 words, and only grammar and spelling mistakes were corrected as data cleaning. The texts differed in their topics (subject areas), and included TEFL/ELT, first/second language acquisition, discourse analysis, corpus-based studies, linguistics, sociolinguistics, and cognitive linguistics.

Nasseri and Thompson examined one lexical density measure and 14 lexical diversity indices and categorized them into six pairs or groups, depending on their similar quantifications (see Table 2.2). *MSTTR, MATTR, HD-D,* and *MTLD* scores were computed using TAALED (Kyle, 2018), and *Vocd-D* scores were calculated using Coh-Metrix (Graesser et al., 2004). The lexical density scores and other nine LD indices were computed using the Lexical Complexity Analyzer (LCA; Lu, 2012). Tokenization, tagging, and lemmatization were carefully considered to ensure reliable comparisons between the scores calculated with the three analysis tools. Since TAALED and LCA were originally based on lemma counts, the data were lemmatized to calculate *Vocd-D* scores in Coh-Metrix.

Table 2.2

Pairs and Groups of LD Measures with Similar Quantification Methods

Category	Measures
1	Number of Different Words-type I (NDWERZ) Number of Different Words-type II (NDWERZ)
2	Mean Segmental TTR (MSTTR; Johnson, 1944) Moving Average TTR (MATTR; Covington & McFall, 2010) Measure of Textual Lexical Diversity (MTLD; McCarthy, 2005)
3	Bilogarithmic TTR (LOGTTR; Herdan, 1960) Uber's U (UBER; Dugast, 1978)
4	Vocd-D (Malvern et al., 2004) Hypergeometric Distribution (HD-D; McCarthy & Jarvis, 2007)
5	Verb Variation-type I (VV1) Verb Variation-type II (VV2) Corrected Verb Variation I (CVV1)
6	Lexical Variation (LV) Noun Variation (NV)

The authors explored (i) the differences in using lexical units (nouns, verbs, types, and tokens) between English L1, ESL, and EFL writer groups; (ii) the correlations among lexical density and diversity indices; (iii) the between-group differences in lexical density and lexical diversity; and (iv) the stronger LD measures to predict writing proficiency. To achieve these four research objectives, the LD scores were analyzed using a bootstrapping method, a general linear model, a one-way ANOVA, and Pearson correlation tests.

First, regarding the lexical profile differences between the three different learner groups, the findings indicated that the EFL writer group used fewer word types and lexical types than the ESL and L1 English writers. Second, the correlational analyses indicated that lexical density was poorly correlated with most LD measures

but moderately and negatively correlated with lexical variation. The correlations between the measures in each category were moderate to high, except for the lexical variation and noun variation indices in category 6. Verb-based indices in category 5 had moderate to high correlations between each other while the correlations were strong for categories 1, 2, 3, and 4.

Third, ten out of the 15 measures showed that the EFL group was significantly different in its lexical density and diversity use from the ESL and L1 English groups. The texts of the English L1 and ESL writers were found to use similar amounts of dense and diverse vocabulary, while the EFL writing texts were the least lexically dense and diverse. However, the verb-based, lexical variation, and noun variation indices could not discriminate between the English L1 and the two other L2 writer groups, showing the similar use of diverse nouns and verbs.

Fourth, among all measures, the lexical density measure could discriminate between the EFL and the other two learner groups with medium effect sizes. However, all nine LD measures were discriminative of all learner groups with the larger effect sizes. Based on the effect size values, the *NDWESZ, UBER, MSTTR, MATTR, HD-D* measures appeared more powerful discriminators. With the highest Cohen's *d* values, *MSTTR* and *HD-D* were the strongest predictors.

To conclude, Nasseri and Thompson's study highlighted that, among the three learner groups, the abstracts written by EFL learners were the least lexically dense and diverse. The lexical density and diversification of both the English L1 writers and the ESL writers who were studying in an immersion program in an English-speaking country, the UK, appear the same. The lexical density and diversity measures used were predictive of academic writing differences, and *MSTTR* and *HD-D* were the

strongest discriminators between the writing proficiencies of the English L1, EFL, and ESL learners.

Nasseri and Thompson's study is a significant study for two main reasons. First, their study appears to be one of the early attempts to give insight into the effectiveness of LD measures which are based on similar calculation methods. There has been rich research information on LD measure predictions on the basis of a single measure or comparison between different measures, irrespective of their similar or different quantifications. However, this study's examination of a combination of different LD measures that share similar calculation methods provides valuable new insights that add to the current LD knowledge.

Second, the authors carefully considered the learner group, particularly L2 learners of English, by differentiating between ESL and EFL learner groups. Mostly, EFL and ESL learners have been lumped together and referred to as "L2 learners" in studies without differentiating between them. Possible reasons for this might be the low data availability or test formats (e.g., the IELTS test targeting L2 learners from diverse contexts, such as different countries and English learning backgrounds). However, EFL and ESL learners might be different in terms of language and lexical knowledge because of their different levels of English language exposure. Nasseri and Thompson carefully considered these two distinct learner groups, and their study indicated that ESL writers' vocabulary (density and diversity) knowledge is indeed higher than that of EFL writers.

Despite these two strengths, the study also includes at least two weaknesses that should be addressed in future LD studies. First, itself addressing a weakness in several of the other studies examined in this literature review, the study clearly stated the specific lexical unit (the lemma count) used to count different words deployed in

the abstracts and showed the predictions of academic writing of various LD measures either as a single index or as combined measures, based on the lemma count. Thus, the study has contributed to the recent argument that the appropriate choice of word counting unit is important for different learner groups. However, the inclusion of some alternative word counting units that can encompass different inflectional and derivational knowledge, such as simple, flemma, or word-family counts might have indicated whether LD measures predictions is dependent on the analysis units deployed.

Second, because of the aim of analyzing lexical profile (e.g., token) differences between English L1, ESL, and EFL writers, the examination of abstracts of varying length (175 – 300 words) seems reasonable. However, maintaining a constant text length is essential in comparing LD scores, so the texts should all be of the same length, as in Treffers-Daller et al. (2018). The different LD measures have been shown to have different levels of sensitivity to text length. For instance, the number of different words (types) is greatly influenced by text sample size, whereas *MTLD* seems less sensitive to text length. Future studies could compare LD measure predictions based on both the same and differing text lengths to glean more insights.

To conclude, Nasseri and Thompson provided new findings about various combined LD measure predictions of writing proficiency by carefully considering and consolidating their calculation methods. Moreover, the authors emphasized the lemma count's effects on LD measures' predictability. However, the study might have made even more of a contribution to the LD research if it had analyzed other word counting criteria and controlled the effects of text length on LD measure predictions.

2.3.10 Zenker and Kyle (2021): Investigating minimum text lengths for lexical diversity indices

Besides providing ample evidence of LD measures' predictive validity, numerous studies (Jarvis, 2007; Koizumi, 2012; Koizumi & In'nami, 2012; Malvern & Richards, 2002; McCarthy & Jarvis, 2010) have attempted to ensure the internal validity of LD measures by determining whether LD measures are reliant on text sample size. These studies have found that text length indeed affected LD measures to varying degrees. However, only a few studies have explored how long the texts should minimally be for each LD measure. To respond to this apparent research gap, Zenker and Kyle (2021) investigated the text length effects of a variety of LD measures and the minimum text length needed for each measure to generate reliable LD scores.

The authors analyzed a written corpus of 4,542 argumentative essays produced by university students in 10 Asian countries: China, Hong Kong, Indonesia, Japan, Korea, Pakistan, the Philippines, Singapore, Taiwan, and Thailand. The participants responded to one of two writing prompts: "*It is important for college students to have a part-time job*" (2,214 essays) or "*Smoking should be completely banned at all the restaurants in the country*" (2,328 essays). The participants' levels were classified at A2, B1, and B2 levels of the Common European Framework of Reference based on the scores of the L2 vocabulary size and for the TOEFL or TOEIC proficiency tests.

For data preparation processing, first, the texts with fewer than 200 tokens were removed. Then, the essays were divided into texts of different lengths: texts of 200 tokens, segments with varying lengths (50 tokens to 200 tokens), and texts with different lengths increasing by 5 tokens. LD scores (*TTR, Root TTR, Log TTR, Maas, MATTR, HD-D, MTLD, MTLD-MA-BI,* and *MTLD-MA-Wrap*) were computed with a

freely available text analysis tool (Kyle, Crossley, & Jarvis, 2021), and the obtained scores for the texts of the same lengths were averaged. As a result, each LD index generated 31 scores for each essay.

Zenker and Kyle analyzed the data with four quantitative analyses detailed below. The raw LD score and z-scores (the relationship between a score and a group's mean score) analyses were conducted to investigate the lengths at which LD measures were stable. Pearson analyses were performed to explore the correlations between LD measures and text lengths, and linear mixed-effect models were used to examine proficiency level effects on LD measures.

First, raw LD values were analyzed to visually indicate the text lengths at which each LD index indicated stability. Line graphs showed that *TTR*, *Root TTR, Log TTR,* and *Maas* indicated no stability for any text lengths, whereas *MATTR* and *HDD* indicated stability right from the beginning, and *MTLD, MTLD-MA-BI,* and *MTLD-MA-Wrap* were stable at 100 tokens. Among all LD measures, *MATTR* was the most stable index across all text lengths.

Second, z-score analysis, which put all LD scores on the same scale, was then conducted to facilitate easier comparisons between different LD indices. The analysis showed similar results to the raw value analysis in that *MATTR, HDD, MTLD* and *MTLD-MA-Wrap* could produce more stable values than other measures. For the traditional measures, *TTR, Root TTR,* and *Log TTR* were not stable at any number of tokens while the stabilizations occurred for *Maas, HDD,* and *MTLD-MA-BI* at 170, 130, and 140 tokens, respectively, and for *MTLD* and *MTLD-MA-Wrap* at 70 and 75 tokens. Of all measures, *MATTR* could hold stable for the whole range of 50-200 tokens.

Third, Pearson analysis was used to explore the extent to which LD measures and text length were related. The findings illustrated the stability of the *MATTR, HDD, MTLD,* and *MTLD-MA-Wrap* across 31 text lengths. Additional correlational analysis with three data bins (50-95, 100-145 and 150-195 tokens) indicated that *MATTR, HDD,* and *MTLD* indices were consistent across all three bins, that *Maas, MTLD-MA-BI,* and *MTLD-MA-Wrap* were stable at the second and third bins, and that *TTR, Root TTR,* and *Log TTR* showed no stability in all bins.

Fourth, linear mixed-effects and visual analyses with line graphs were applied to investigate how the participants' proficiency levels influenced LD measures. Applying various LD measures was not appropriate for this analysis, so Zenker and Kyle selected a single measure, *MATTR*, which was the most stable measure. The analysis proved the small yet significant effect of proficiency on the LD measure. The participants at higher levels used greater numbers of different words than the participants at lower levels. For instance, B1 level texts were more lexically diverse than A2 level texts.

In summary, Zenker and Kyle showed that *MATTR, HDD,* and *MTLD* were stable measures for texts of 50–200 tokens, *Maas, MTLD-MA-BI,* and *MTLD-MA-Wrap* were stable for texts of 100 or more tokens, and the traditional measures (*TTR, Root TTR, Log TTR)* did not stabilize at any text length. Moreover, the researchers found that the participants' proficiency level did affect the LD measures.

Zenker and Kyle's paper is important for three strengths: (i) new findings of the minimum text length required for each LD measure; (ii) the statistical advantages of different analyses for more reliable findings; and (iii) the use of a large corpus with participants from a wide range of Asian countries.

First, the paper appears to be the first attempt to reveal how various LD indices are influenced by different text lengths in a single study, and it surveyed and suggested the specific text lengths needed for each index to produce valid LD scores. The study identified *MATTR* as the most stable measure for short texts of 50–200 tokens among all nine LD measures under study and added evidence of the traditional measures' reliance on text length. The study's important findings can guide LD researchers to undertake careful LD measure selection, depending on the particular text lengths used in their studies, to generate more valid findings and conclusions.

Second, the authors confirmed the findings of LD measures' dependency on text length by using different statistical analyses on both raw LD scores and z-scores. As all these analyses yielded consistent findings, the conclusions and generalizations drawn gain greater validity.

Third, Zenker and Kyle analyzed a large corpus which included L2 learners from different Asian countries. Such large participant sample size analysis seems more representative of the general population and provides accurate statistics, enabling more reliable findings. Specifically, the study's findings seem highly generalizable to the L2 English Asian context.

Despite these three strengths, the paper has at least three weaknesses: (i) the lack of requisite data treatment for LD assessment; (ii) the unclear information on the word counting technique; and (iii) the exploration of the potential effect of participants' L1 backgrounds on the stability of the LD measures.

First, Zenker and Kyle's study lacks information about the data cleaning which is necessary to generate reliable LD scores. Removing words that do not show the participants' lexical diversity knowledge (e.g., proper names, acronyms, cardinal numbers) prevents LD score inflation (Treffers-Daller et al., 2018). The reported

findings would be more reliable if the authors could have carefully treated the data prior to LD score calculation. To build on the strength of existing LD studies, future researchers should follow or base their data treatment procedures on those tried and trusted ones exemplified in previously published studies.

Second, Zenker and his colleague did not clearly indicate how they counted different words, i.e., they did not explain whether they used simple type, lemma, or word family counts. Researchers (Treffers-Daller, 2013; Treffers-Daller et al., 2018) assert that the conceptualization of what constitutes a different word (type) is important in LD assessment since the use of different word counting units (lemma, flemma, word-family counts) might affect LD measures and scores. Selecting the appropriate word unit for analysis should be based on the participants' inflectional and derivational word knowledge. For instance, a word-family count might not be suitable for L2 learners with inadequate derivational knowledge (McLean, 2017; Brown et al., 2020). It would have been more informative if the authors could have mentioned how they defined and counted different words. Future research should focus on the appropriate selection of the word counting unit that neither underestimates nor overestimates the participants' existing lexical knowledge in order to generate more valid and reliable LD findings.

Third, despite the new insights into the minimum text length required for various LD measures to be reliable, the findings are applicable to only Asian learners of L2 English from different nations. Since L1 background is considered one of the key factors affecting the L2 learners' diverse vocabulary use and affecting the LD measures predictions (Yu, 2010), it is well worth examining whether the reported findings of the LD measures' stability will be the same and applicable for each specific L1 background.

To conclude, Zenker and Kyle's study indicated the minimum text lengths needed to ensure the stability of nine different LD indices in assessing the diverse vocabulary of Asian L2 English learners by confirming the findings with different statistical analyses, indicating that *MATTR* is least affected by text lengths (50–200 tokens). However, the study pays scant attention to data cleaning processes, the influences of word unit selection, and learner first language background, even though these play essential roles in LD assessment.

2.4 Analysis unit selection in L2 vocabulary assessment

As the review of LD studies (section 2.3) has indicated, the importance of the analysis unit choice in LD assessment had been a long-overlooked factor that has now become the subject of both growing attention and contention. Thus, this section discusses the recent controversial issue in L2 vocabulary assessment: Is one analysis unit better than the other units? This section includes two sub-sections. The first sub-section explains how the analysis unit choice is important in assessing L2 vocabulary knowledge by presenting various perspectives on different lexical units' appropriateness to particular contexts, referring to recent key studies (Brown, 2018; Brown et al., 2020; Brown et al., 2021; Dang, 2021; Kremmel, 2021; Laufer, 2021; Mclean, 2017; Nation, 2021; Stoeckel et al., 2020; Webb, 2021). The second sub-section highlights the important research gap that might be addressed to help deepen if not resolve this debate: the need to explore the flemma count, which has as yet remained unexplored in LD assessment. This sub-section reviews four important papers, which have raised the awareness of the importance of the careful analysis unit selection to suit particular L2 learners and which have questioned the appropriateness of the conventional analysis unit (i.e., the word-family count) for L2 learners with insufficient derivational knowledge.

2.4.1 Analysis units reflecting learners' lexical knowledge

There has been a growing interest in the choice and comparisons of the analysis units and how they can reflect L2 learners' lexical knowledge, emerging in contrasting views on the choice between word types (simple count), lemma, flemma, or word-family counts. Nation (2021) clarified that "the primary issue behind the word families debate is learner knowledge" (p. 969). Different analysis units reflect learners' different word part (inflections and derivations) knowledge levels (see Table 2.3). L2 vocabulary researchers have long been referring to Bauer and Nation's (1993) word-family levels although the scheme was created using the criteria of morphological factors (e.g., frequency, productivity) rather than being mainly based on learner knowledge (Nation, 2021).

Table 2.3

Teaching Order of L2 English Derivational Affixes (Bauer & Nation, 1993)

Level 1	A different form is a different word.
Level 2	Inflectional categories: plural -s, past tense -ed, comparative -er, etc.
Level 3	The most frequent and regular derivational affixes: -able, -er, -ish, -less, -ly, -ness, -th (fourth), -y, non-, un-
Level 4	Frequent and regular affixes, e.g., -al, -ation, -ess, -ful, -ism, -ist, -ity, -ize, -ment, -ous, in-
Level 5	Infrequent but regular affixes, e.g., -age -ance, -ship, mis-, etc.
Level 6	Frequent but irregular affixes, e.g., -ee, -ic, -ion, re-, etc.
Level 7	Classical roots and affixes, e.g., -ate, -ure, etc.

Note. Table was adapted from Leontjev et al. (2016).

Four common analysis units are the word type (simple), lemma, flemma, and word-family counts. Word type considers the different inflectional and derivational

forms of the same words as different types. Orthographically identical forms, such as "bank" (the financial institution) and "bank" (the edge of a body of water), are counted as the same type, whereas orthographically distinct forms like "see" and "sees" are counted as separate types. A lemma count considers a base word and its inflected forms under the same part of speech as the same type, while a flemma count considers a base word and its inflected forms as the same type irrespective of the parts of speech. Thus, the words, “developed” (verb) and “developed” (adjective) are different lemmas but the same flemma. For a word-family count, a base word and both its inflected and derived forms are considered the same type, so “develop, develops, developed, developing, developer, development” all constitute the same type.

Recently, several studies (Brown, 2018; Brown et al., 2020; Mclean, 2017; Stoeckel et al., 2020) have provided important information relating to L2 learners’ existing inflectional and derivational knowledge and have raised awareness regarding the criticality of analysis unit choice in L2 vocabulary assessment. Brown (2018) highlighted the complexity of word lists (e.g., word forms with multiple affixes, challenging affixes, infrequent base words) for L2 learners with inadequate word family knowledge. Brown et al.’s (2020) recent brief review of the empirical studies suggested that L2 learners may be able to demonstrate lemma knowledge receptively, but their derivational knowledge was limited. Stoeckel et al. (2020) also indicated the challenges of low-proficiency Japanese EFL learners with inflections under different word classes and thus proposed the preferability of using a lemma count over a flemma count for such learners. Such studies have concluded that the word-family count might overestimate the vocabulary knowledge of most L2 learners with limited derivational knowledge, indicating that either lemma or flemma counts, which require

only inflectional knowledge, might be a better unit for more accurately gauging the LD of such learners.

As a response to their claim, Webb (2021) argued against these studies' overgeneralization of the word-family count's unsuitability for most L2 learners and contended that it should not be assumed that just one unit best fits all learners regardless of the diversity of their levels and backgrounds. He thus suggested that the smaller analysis units (flemma and lemma counts) might be more suitable for low-proficiency learners, who seem unable to consider different forms of the words, whereas the larger unit (word-family count) might be more appropriate for high-proficiency learners, who have sufficient inflectional and derivational knowledge. Webb also stated that the analysis unit choice is likely to be dependent on several factors, including vocabulary size, morphological knowledge, and/or proficiency. Similarly, Dang (2021) agreed with Webb by implying that the word unit choice in the wordlist creation should match the list's purpose.

In their recent study, Brown et al. (2021) highlighted that it is unreasonable to consider just one particular analysis unit to be the most suitable for all L2 learners. They also outlined similar ideas to Webb (2021) that the smaller lexical units might possibly be better for low-level learners, whereas the larger units might be more suitable for advanced learners. However, Kremmel (2021) has argued for careful consideration in differentiating between the use of the smaller units (lemma or flemma counts) and the larger unit (word-family).

Despite the invaluable longstanding contributions of the word-family count to the L2 vocabulary assessment field, Brown et al. (2021) pointed out that "the use of the word family remains common not just in areas where its use is well founded but also in others where it is not" (p. 952). They indicated that examining the

appropriateness of the word-family count use appears a necessity because of how prevalently it has been used, and it is thus a useful starting point for a possible paradigm shift. Similarly, Kremmel (2021) called for more experimental and replication studies on the various analysis units so as to make valid conclusions based on the richer evidence. Nation (2021) reminded vocabulary researchers to interpret the results with caution if the analysis unit used and the participants' proficiency are not well-matched.

2.4.2 Studies suggesting alternative units (lemma or flemma counts) to the word-family unit in L2 vocabulary assessment

This section reviews four recent studies (McLean, 2017; Brown et al., 2018; Brown et al., 2022; Stoeckel et al., 2020) which reported evidence of the often-lacking word part knowledge of L2 learners and thus proposed the use of lemma or flemma counts instead of a larger unit (word-family count). This section discusses how these studies highlighted a research area that should not be ignored in L2 context in relation to the analysis unit. I also explain why it might be of great value to examine the flemma count's usability in the L2 LD assessment field, based on the doubts about the word-family count's suitability for L2 learners with insufficient derivational knowledge.

2.4.2.1 McLean (2017): Evidence for the adoption of the flemma as appropriate word counting unit

The word-family (referred to as WF6; Bauer & Nation; 1993), which includes the base word and its inflections and derivations, is a widely used word-counting unit in L2 vocabulary assessment. However, some doubts arise about the word-family unit appropriateness for L2 English learners because of their limited derivational

knowledge. More empirical evidence on L2 learners' actual inflectional and derivational knowledge is needed to confirm WF6's presumed overestimation of L2 learners' lexical knowledge, but some has already emerged.

McLean (2017) examined the comprehension of the base words and related inflected and derived forms of L1 Japanese L2 English learners. The participants were 279 Japanese university students who belonged to their institution's intermediate, advanced, and upper-advanced English levels. They were 235 first-year students, 21 second-year students, 10 third-year, and 13 fourth-year English major students with one year experience of studying abroad. McLean divided the participants into three lexical proficiency groups: beginner (n=85), intermediate (n=177), and advanced (n=17) groups based on the New Vocabulary Levels Test (NVLT; McLean & Kramer, 2015) scores. During the 30-minute test, the participants completed 24 multiple-choice items per 1,000-BNC/COCA WF6 band, for each of the first five 1000-word bands, by selecting the correct Japanese translation from the four given options for the target word.

McLean also investigated the participants' inflectional and derivational knowledge of the base words with a 100-item comprehension test. The test consisted of twelve known target word families (*use, move, collect, center, teach, accept, maintain, develop, standard, circle, adjust,* and *publish*), featuring with 87 inflected and derived forms (e.g., -ed, -ing, -er, -less, -ized, non-, sub-, re-able). The author excluded the two inflected forms of plurals (-s) and third person singulars (-s, -es) because of their negligible effects on reading comprehension, and randomly arranged the target words to prevent the guessing effects. Three L1 Japanese teachers of L2 English assessed the test responses: one main teacher marked all participants' responses, and the other two teachers marked 20% of the participants' responses,

which had been randomly selected. The evaluators did not differentiate between word classes: for instance, if the participants translated "*center*" as either a noun or a verb form, they marked both answers correct.

Prior to analysis, McLean checked whether the participants knew the meanings of the target words. If a participant did not know the meaning of a word, the participant's responses to the base, inflected, and derived words were excluded. He conducted three analyses to explore the comprehension abilities of these L1 Japanese learners of English of the inflected and derived forms of the base words.

First, the author used Cochran's Q analysis to explore whether there was a significant difference in understanding of the base, inflected, and derived forms. The analysis showed that the differences between participant knowledge of the base words and related inflected and derived forms were statistically significant for all target word families across all three groups with large effect sizes. The findings suggested that participant base word knowledge did not correlate with the inflectional and derivational knowledge of that word. The author highlighted the inappropriateness of word family unit in assessing receptive vocabulary knowledge (comprehension) of L2 English learners.

Second, McLean conducted the additional Cochran's Q analysis to investigate whether the participants' knowledge of the base words and associated inflected words were different. The results showed that significant differences with small effect sizes were found for only three target words (*center, develop,* and *circle*) among the beginner and intermediate groups. However, the advanced group did not indicate any significant differences at all. The findings highlighted that the participants could indeed understand the inflected forms if they knew the base words, thus suggesting

the flemma count's suitability as a written receptive word-counting unit for L1 Japanese L2 English learners.

Third, the author conducted McNemar Chi-square test analysis to further investigate the suitability of adopting the flemma count as a word-counting unit. The analysis illustrated that a flemma count underestimated all participants' knowledge of three derivations (*user, teacher, publisher*). However, beginner and intermediate groups indicated significant differences with small effect sizes in understanding the base words and related inflections for *center, circle,* and *develop*. For the advanced group, the participants could comprehend the inflectional forms but manifested limited derivational knowledge, as 32 out of 51 derivatives seemed difficult for them. The author therefore proposed the flemma count as more appropriate than WF6 for the participants under study.

To conclude, the study showed that L1 Japanese L2 English learners with low lexical proficiency (i.e., beginning and intermediate levels) could comprehend the inflectional forms but not the derivational forms. Moreover, the advanced level participants had sufficient inflectional knowledge but limited derivational knowledge, even though this finding was based on a small participant sample size. The study's findings were in line with Ward and Chuenjundaeng (2009), who had also identified the poor relationship between L2 learners' knowledge of the base words and their derivations. The author thus confirmed that the word-family count seems challenging for L2 learners such as those under investigation, and that the flemma count should instead be adopted as a more suitable L2 written receptive vocabulary measurement unit.

McLean's study is significant as it is one of the few empirical studies to explore L2 learners' actual inflectional and derivational knowledge for receptive

purposes when the L2 vocabulary research mainly relies on word-family based wordlists. The study highlighted the word-family count's inappropriateness in some L2 contexts and instead proposed the suitability of a lemma or flemma count to more accurately assess L2 receptive vocabulary knowledge. The study therefore highlighted the need to reevaluate the word-family count's suitability in assessing the vocabulary knowledge of learners with low lexical proficiency, and it drew attention to the importance of carefully selecting appropriate lexical units that accurately reflect L2 learners' existing lexical knowledge.

Despite its significance, the study has at least three limitations that need further research. First, the author examined a few target word families, so the findings' generalizability seems restricted to just those target words. Further studies should therefore conduct a larger study with a greater number of word families to provide stronger evidence of L2 learners' existing inflectional and derivational knowledge.

Second, the assessors did not discriminate between lemma and flemma while evaluating the test scores. In reality, lemma and flemma counts represent different levels (Levels 2 and 2.5, respectively) in Bauer and Nation's scheme, with the former requiring knowledge of inflections under the same word class, whereas the latter demands knowledge of inflections that belong to different word classes. It would provide clearer information on the better unit (flemma or lemma count) in L2 vocabulary assessment if future research can separately examine these two analysis units.

Third, the study proved that the word-family count is challenging for L2 learners, so the flemma count is a more appropriate written receptive vocabulary assessment unit (meaning comprehension). However, an even greater contribution to

the existing L2 vocabulary literature would be if future studies could explore whether they should also apply the flemma count as the most suitable analysis unit of productive vocabulary in the L2 context.

To conclude, McLean indicated that low-proficiency L2 learners have sufficient inflectional knowledge but limited derivational knowledge, and he thus contended that word family is beyond most L2 learners' existing word knowledge. He therefore indicated greater suitability of the flemma count than the word-family count in L2 settings. However, the lack of discrimination between the lemma and the flemma counts renders the information unclear as to which unit (lemma or flemma count) better suits the accurate representation of L2 learners' lexical knowledge. Moreover, the study solely focuses on receptive vocabulary knowledge, so it leaves an important gap for the examination of the most suitable word units for accessing L2 productive vocabulary knowledge.

2.4.2.2 Brown (2018): Examining the word family through word lists

The available empirical evidence that the conventional word-family count is actually often beyond L2 learners' vocabulary knowledge draws attention to the wider evaluation of word-family based word lists that L2 vocabulary research, pedagogy, and curriculum heavily rely on. Therova (2020) indicated that understanding the characteristics of the currently available word lists is a necessity for more effective use. Brown (2018) analyzed the most common word-family based lists, the Nation's (2006) British National Corpus-based word lists. Brown identified the lists' characteristics and the coverage level that the lists can provide for L2 learners who are unable to consider word family.

The BNC lists include not only the word families of Bauer and Nation's (1993) scheme but also additional forms beyond the scheme (i.e., irregular verb and

noun forms, abbreviations, compound words, and alternative spellings of the base words). Brown examined the first five frequency bands of the lists, which represent a great number of the words in most texts and thus aim to facilitate efficient L2 vocabulary learning, as suggested by Webb and Sasao (2013).

For the analysis, Brown selected 100 word families from each band by using systematic random sampling and calculated the frequency of the 2,396 word forms found in the five samples of 100 word families. He conducted two analyses to explore how and to what extent the lists posed challenges, in terms of word family and the text coverage levels, for learners with different word family knowledge levels (e.g., text coverage level for learners with Level 2 affix knowledge).

To analyze the BNC word lists' characteristics (size and complexity), the author calculated: (i) the number of word forms in the families for all bands; (ii) the number of word forms with different numbers of affixes (e.g., single, two, or multiple affixes) for each band; (iii) the number of word forms in each band that represented the different levels of affix knowledge in Bauer and Nation's scheme; and (iv) the number of word families with infrequent head words.

First, the analysis indicated that the 5000 word families of the BNC word lists comprised around 25,000 word forms, including word forms with zero affixes (e.g., irregular verb and noun forms, abbreviations), and the 1K band comprised 6,000 word forms. Second, regarding the number of affixes in word forms, the findings revealed single affixes in most forms, multiple affixes in 5,000 forms, and multiple derivational affixes in 2,000 forms. The 1K and 2K bands included more forms with two or three affixes than the 3K, 4K, and 5K bands. Third, for the number of word forms representing different levels of Bauer and Nation's scheme, the findings illustrated that Level 2 affixes made up two-thirds of the word forms, over 50% of the

1K forms, and 70% of the 3K, 4K, and 5K bands. Brown also pointed out that Level 2 affixes were not necessarily inflections and did not follow any systematic functional or grammatical rules. Fourth, for one-fifth of the word families, their base words were not frequent.

Brown performed additional analysis to determine the text coverage levels for learners with insufficient derivational knowledge. He estimated the coverage levels based on the proportions of the occurrences of the word families at Bauer and Nation's different word-family levels, plus the additional forms. The findings indicated that, on average, the base word made up 62% of the occurrences of all members in a word family, level 2 accounted for 23% and other levels accounted for 2-5%. Affix knowledge at different levels influenced the actual text coverage. For instance, for 95% coverage level texts, the actual text coverage levels were 58.9% for learners with base word knowledge alone, 60.1% for learners with base word and additional form knowledge, 82.3% for learners with knowledge of base word, additional form, and level 2 affixes.

To conclude, demonstrated the complex characteristics of the first five bands of BNC word lists and the resultant varying text coverage levels that the lists brought to learners with different affixation knowledge. He highlighted that the 1K band, which is supposed to serve as a starting point for learners, paradoxically seemed more difficult than assumed as it contained a greater number of word forms, word families with multiple affixes, and more challenging affixes. Overall, the author proved the oversights and challenges of the word-family count, and implied that lemma or flemma count might instead be more suitable lexical units for accurately evaluating the LD of L2 learners.

Brown's study is significant as it is one of the few studies that identify BNC word lists' characteristics that emphasize word families, thus providing two important insights: the challenges of BNC word lists, and the influence of affix knowledge on text coverage.

First, Brown's detailed examination of the lists reveals the underlying complexity of the lists and provides useful information for the current debate on how to accurately gauge the true extent of L2 learners' word family knowledge. The study showed that the lists contain challenging word forms, in particular that the 1K band is more challenging than it is claimed to be, and that word family is difficult for L2 learners. These findings are crucial for research and pedagogy and also raise awareness of the need for careful consideration of the analysis unit for the word list development or creations. For instance, it is important to decide whether the lists should be word family- or lemma-based if researchers are targeting L2 learners since word-family based lists solely based on L1 corpora may not be appropriate for L2 learners.

Second, the study clearly explained how affix knowledge influenced text coverage levels. Brown estimated the actual text coverage levels for learners with affix knowledge at all six different levels of Bauer and Nation's scale. The assumed coverage level might well be lower if learners cannot deal with word families. The findings are highly beneficial for researchers examining vocabulary coverage, size, and comprehension and the most suitable word unit for L2 learners.

Despite these significances, the study includes one main limitation relating to the analysis unit: the text coverage level for learners with Level 2.5 affix (flemma) knowledge. The study examines all six different affix levels including other forms beyond the scale but does not explore the flemma count. The flemma count is also a

common word-counting unit, but it should be separated from the lemma count because of the flemma count's need for learners' word class knowledge in considering the inflectional knowledge.

Recently, growing concerns about the inappropriateness of the word family count in the L2 vocabulary context have led to proposing the use of lemma or flemma counts. However, since a lemma count underestimates proficient L2 learners' inflectional knowledge (McLean, 2017), the flemma count has gained interest in L2 vocabulary research. Therefore, it might have been more informative if Brown had determined the text coverage level for learners with inflectional knowledge beyond word classes.

To conclude, Brown showed the complexity of the BNC word lists with regards to the word family; and the challenges of the 1K band, as well as the different text coverage levels for learners with different affix knowledge levels. However, the study could have provided more insight into the analysis unit selection by considering the flemma count, which requires learners' higher word class knowledge than a lemma count.

2.4.2.3 Brown et al. (2022): The most appropriate lexical unit for L2 vocabulary research and pedagogy: A brief review of the evidence

The selection of the appropriate lexical unit in vocabulary assessment influences the validity of a study's findings or conclusions. The word family, which comprises a wide range of inflected and derived word forms, has been the dominant lexical unit in both L1 and L2 vocabulary research for years. Recent studies (Brown, 2018; McLean, 2017; Schmitt and Zimmerman, 2002) have analyzed various lexical units in different contexts (e.g., different learner groups, proficiency levels or tests), and have discovered that the word-family count is challenging for L2 learners. To

understand the current knowledge of L2 learners' ability to consider word families and to guide future research directions, summarizing the empirical findings of the existing literature is essential. Brown et al. (2022), therefore, briefly reviewed some empirical studies exploring which lexical unit is a more appropriate assessment unit of learners' receptive vocabulary and the effects of the word unit choice on text comprehension.

Brown and his colleagues addressed two questions: (i) whether L2 English learners could recognize the relationships between the different base words; and (ii) how much affixation knowledge was needed for reading comprehension. To address their first question, the authors reviewed and summarized the findings of six studies exploring which lexical unit best matched L2 learners' receptive vocabulary knowledge to represent it accurately. The first study was Stoeckel et al. (2018), who explored L1 Japanese L2 English learners' understanding of different word classes by using a translation test in which they presented the target words in sentences. Brown et al. summarized the study's findings and found that word knowledge in one word class related to other word classes just over half of the time (56%), suggesting the lemma count as a more suitable unit than the flemma count.

The second study was Mochizuki and Aizawa (2000), who examined L1 Japanese L2 English learners' acquisition of derivational affixes with a multiple-choice test. Brown et al. highlighted the study's findings that learners could identify 56% of the prefixes and 67% of the suffixes, and that affixation knowledge increased with the lexically more proficient learners; however, even advanced learners showed incomplete knowledge of common affixes.

The third study was Sasao and Webb (2017), who explored whether L2 learners with multiple L1 backgrounds (not specified) were able to recognize the

form, meaning, and grammatical functions of the affixes by using a multiple-choice test. The study's findings indicated 73% for form recognition, 83.7% for meaning recognition, and for recognition of 62.9% for grammatical functions, implying that learners had some knowledge of affixes. However, Brown et al. noted that Sasao and Webb did not explain the learners' proficiency levels.

The fourth reviewed study was Ward and Chuenjundaeng (2009), who investigated L1 Thai L2 English learners' affix knowledge by having learners translate the base words and the derivations with four frequent suffixes. Brown et al. summarized the study's findings that learners' knowledge of one part of speech (POS) knowledge did not equate to the comprehension of other POS.

The fifth study was McLean (2017), which examined L1 Japanese L2 English learners' capability to understand different POS by using a translation test. The study highlighted significant differences between knowledge of the base words and the related inflected and derived forms for beginner, intermediate, and advanced learners.

The sixth study that Brown et al. analyzed was Kremmel and Schmitt (2016), who examined English L1 Australian late secondary school learners to investigate the relationships between the recognition of the base words and their meanings. The study's findings highlighted weak correspondence between comprehension of the base word meanings and those of the derivational forms.

To address the second question, Brown and his colleagues reviewed two studies designed to explore how the word unit choice influenced reading comprehension. The first study was Laufer and Cobb (2019), who examined 21 English texts across four genres to investigate the amount of the affix knowledge necessary for reading comprehension. The study showed that the proportion of the

derivational forms ranged from 3.17% for six graded readers to 7.78% for five academic papers, and the average proportion was 5.6% of the total words used.

The second study was Brown (2018), who conducted a corpus-based examination of the derivational knowledge influence on text coverage. The investigation of the first 1000-word family band of Nation's British National Corpus (BNC) word lists showed derivational forms constituted 13.4% of the total frequency of the family, implying that, with 95% text coverage provided by the first 5000 word families, the learners knew only 82.3% of the total words if they lacked the derivational knowledge.

To conclude, Brown et al.'s brief review showed that the word-family unit seemed beyond the lexical knowledge of L2 learners from different L1 backgrounds (i.e., Japanese, Thai), and no single study supported the word-family count use in the L2 context. Moreover, the authors concluded that even a small proportion of derivations in a text might influence learners' comprehension and thus questioned Laufer and Cobb's suggestion that learners' knowledge of few affixes might not have hindered their comprehension. Based on the evidence of the reviewed studies, Brown et al recommended reassessment of the suitability and usability of the conventional word-family count in L2 vocabulary studies.

Despite being a brief review on a few studies, Brown et al.'s review has contributed sizably to L2 vocabulary research as it collects and merges the scholarly studies and provides insight into the current knowledge and understanding of the appropriate word units for L2 learners. The review presented convincing evidence of L2 learners' difficulty when encountering word families, suggesting the greater suitability of lemma or flemma counts compared to a word-family count. The review showed the effects of analysis unit selection on text comprehension. Future L2

vocabulary studies can therefore carefully select the alternative word unit (lemma or flemma count) for more valid findings.

Despite this significance, the review included at least two limitations that need to be further addressed: the examination of the appropriate analysis unit for assessing productive vocabulary, and detailed investigation of the suitable word-counting units depending on learners' proficiency levels.

First, Brown et al.'s review was limited to empirical studies that examined the suitable lexical unit only for accessing receptive vocabulary knowledge. It might have been more informative if the authors could have included studies examining the appropriate word-counting units for assessing productive vocabulary knowledge, such as Treffers-Daller et al.'s (2018) study that supports the lemma count's use over simple type and word-family counts in lexical diversity assessment. Then, the paper might have yielded richer insight into the most appropriate word unit choices for both receptive and productive vocabulary assessment.

Second, the authors explained the participants' L1 backgrounds (i.e., Japanese, Thai), as well as the test types used in the reviewed studies (e.g., translation test, multiple choice test); however, they paid less attention to the participants' language proficiency levels that might have provided important information on the word unit selection. Learners at different proficiency levels might need different word units. Therefore, for clearer information about the word unit that best fits learners at specific levels, future reviews of studies should carefully consider the language proficiency levels of the participants in the studies being reviewed.

To conclude, Brown et al.'s study explored the ongoing debate concerning the possible overestimation of word-family count on L2 learners' vocabulary knowledge. The study concluded that the word-family count is challenging, so a lemma or flemma

count seems more appropriate in assessing L2 receptive vocabulary knowledge, and that the lexical unit choice influences text comprehension. For deeper insights, though, future reviews of research on word unit suitability should evaluate studies of both receptive and productive vocabulary.

2.4.2.4 Stoeckel et al. (2020): Is the lemma more appropriate than the flemma as a word counting unit?

Selecting an appropriate analysis unit is essential in vocabulary testing as it might affect the study's findings or conclusions. For instance, in lexical diversity assessment, different ways of classifying and counting words can have different impacts on the LD measures predictions of L2 language proficiency, leading to different interpretations of the LD measures' predictive power (e.g., Treffers-Daller et al., 2018; Yu, 2010). The choice and use of word-counting unit should closely match the particular participants' lexical knowledge. For most L2 English learners, the conventional word-family count seems somewhat unsuitable (Brown, 2018), so a lemma count could be more suitable as it can show the learners have knowledge of inflections under the same word class despite limited knowledge about derivative forms of the words (McLean, 2017). However, learners' inflectional knowledge under different parts of speech (POS) remains unknown. Therefore, Stoeckel et al. (2020) extended McLean's study by exploring whether the lemma or flemma count is a more appropriate lexical unit for gauging the L2 learners' lexical diversity knowledge.

The participants were 64 L1 Japanese students from two colleges in Japan and their TOEIC (Test of English for International Communication) and CASEC (Computerized Assessment System for English Communication) scores were used to assign equivalent English proficiency levels at CEFR: A1 (5), A2 (25), B1 (31), and B2 (3). To explore the participants' receptive knowledge of the words representing

different word classes, the authors used a sentence translation task in which the participants needed to translate the whole sentence, not just the target word. The sentences included 12 target words of *edit, result, pause, export, quote, rise, fool, extra, function, twist, variable*, and *compound*. These words came from two different word classes but had the same sense of meaning (e.g., *pause* as a noun and as a verb) and represented the first 3000-word level. The participants could access the sentences only once as the online format and translated or responded "*I don't know*" if they did not know the meaning. They randomly separated the same words to avoid the priming effect. Two L1 Japanese evaluators with SLA/TESOL backgrounds conducted the dichotomous scoring of the participants' knowledge of the target word meanings, and the ratings were reliable and consistent ($\alpha = .87$).

To examine the links between the learners' understanding of words in one POS and in another POS, the authors calculated the Jaccard's index (the value range from 0 to 1), which was the division of the participants into those who had knowledge of both POS for a word and those who had knowledge of only one POS. The findings indicated the Jaccard values ranged from (.00) to (.82), showing that there were variations among the results for the 12 target words. For instance, the participants who knew the word "*edit*" in one POS could comprehend it in another POS (.82), but not the word "*compound*" (.00). The participants seemed to understand the different POS of more common words (e.g., *edit, result, pause, export*) rather than the less common words (e.g., *twist, variable, compound*). The total Jaccard value for all 12 words under study was .56, meaning that the students with lexical knowledge of a word in one POS could comprehend the inflection of the word in another POS only 56% of the time. The Jaccard value should be closer to (1), i.e., indicating the stronger

inflectional knowledge of words to recommend the flemma count's use. Therefore, the authors concluded a lemma count was more appropriate than a flemma count.

To conclude, the authors highlighted the greater suitability of a lemma count than a flemma count since the participants' receptive knowledge of target words in one POS was not fully associated with the comprehension of those words in another POS despite having the same written forms and the same meaning sense.

Stoeckel et al.'s study is significant as it represents one of the earliest studies to compare lemma and flemma counts. Several vocabulary studies have focused on the appropriateness of the most widely used lexical units, such as word type, lemma, or word-family count, and provided insight into the usefulness of each unit. However, Stoeckel and his colleagues investigated the flemma count, which had not yet been explored. They thus discovered the important information that a flemma count should not be used in testing the receptive vocabulary knowledge of L2 English learners at low to intermediate levels due to the poor associations between word knowledge in different POS, leading them to suggest instead the use of a lemma count. Based on the findings, we can conclude that we should apply different analysis units for learners who have different inflectional and derivational knowledge levels.

Despite this contribution, the study left at least two issues that might need further investigation: an examination of the use of a flemma count in analyzing learners with higher language proficiency, and an investigation of the appropriate unit (lemma or flemma count) in evaluating productive word knowledge.

First, the study identified some participants with word knowledge in one POS as encountering difficulties in understanding the word meanings in another POS. However, the participants under study were ESL learners with low to intermediate proficiency levels. The findings are thus not generalizable to higher-proficiency

learners as more proficient learners might understand the words across such POS boundaries. Therefore, future research should examine a wider range of proficiency levels.

Second, Stoeckel et al. determined which word unit was more appropriate in testing the meaning comprehension by comparing the suitability of lemma and flemma counts and then recommending the lemma count as the better analysis unit for the participants' L2 receptive vocabulary assessment. No study to date, however, has compared lemma and flemma counts in assessing productive vocabulary knowledge (i.e., the vocabulary range used in spoken or written products) of L2 English learners. It would be more insightful if further research could discover which analysis unit we should adopt for best accessing the productive vocabulary of L2 learners.

To conclude, Stoeckel et al. recommended the use of a lemma count over a flemma count in assessing the meaning comprehension of low-level learners in the L2 context. Future research should extend their study by investigating whether more advanced level learners with one POS knowledge of the words can also understand the words in other POS, and whether a lemma or flemma count is more suitable for productive vocabulary assessment.

2.5 Discussion

This section provides a summary and overview of the studies reviewed in section 2.3 and their various attempts to incorporate some of these four influential factors (lexical unit, L1 background, L2 proficiency, and text length) into their studies (see Table 2.4). This section also explains how these four factors are each controlled in the four respective experimental chapters by partially replicating Treffers-Daller et al. (2018) in stating the research questions to be investigated by each experiment.

Earlier LD studies have reported LD measure validity, especially their L2 language proficiency predictions; however, multiple factors have also been found to influence LD scores and LD measures predictions. As this literature review has shown, the analysis unit, L1 background, language proficiency, and text length are important factors that can affect LD measures. Table 2.4 shows that, although some reviewed LD studies have addressed one or two of these four important factors, other studies reviewed did not control for any of these factors. In assessing diverse vocabulary of L2 learners of different L1 backgrounds and different L2 language proficiencies, it seems essential to control as many factors as possible in order to gain deeper insight into the LD measures' applicability.

Table 2.4

Reviewed LD Studies and Their Attempts to Address Four Influential Factors

LD study	Analysis unit	L1 background	Language proficiency	Text Length
Read & Nation	Possible use of simple count	Various L1s	-	Varying
Yu	Simple count	Various L1s, but separate examinations of two major L1 groups (Filipino and Chinese)	-	Varying
Jarvis	Lemma count	L1 English, Swedish, and Finnish	-	Varying
McCarthy & Jarvis	Possible use of simple count	L1 English, Swedish, and Finnish	-	Varying
Gonzalez	Possible use of simple count	14 L1s, including English	-	Varying
Jarvis	Possible use of simple count	L1 English, Swedish, and Finnish	-	Varying
Treffers-Daller et al.	Simple, lemma, word-family counts	Various L1s	-	Constant
Zhang & Daller	Possible use of simple count	L1 Chinese	-	Varying
Nasseri & Thompson	Lemma count	L1 English and Iranian but with different L1s	-	Varying
Zenker & Kyle	Possible use of simple count	Various Asian L1s	-	Varying

First, regarding the lexical unit, the analysis unit choice has gained greater prominence in LD assessment. Most LD studies discussed in Section 2.3 may have used the simple count (Jarvis, 2017; McCarthy & Jarvis, 2013; Read & Nation, 2006; Yu, 2010; Zenker & Kyle, 2021; Zhang & Daller, 2019). Yu mentioned his intentional use of the simple count due to the inclusion of few inflections in the data, which would not have influenced the results. However, the other six LD studies

reviewed did not clearly explain the word definition they used, so we might assume the method used to have been that of simply counting different words (i.e., a simple count). Two studies (Jarvis, 2013; Nasseri & Thompson, 2021) examined LD measure predictions of writing proficiency based on a lemma count.

Treffers-Daller et al. (2018) conducted a wider-ranging investigation into the analysis units by comparing the three different analysis units (simple, lemma, and word-family counts). Their findings indicated that these three analysis units influenced LD measures predictions of L2 general language proficiency of the high proficiency learners in different ways. The findings indicated a lemma count could enable and enhance LD measures' predictive power better than either simple or word-family counts.

Thus, with this greater awareness of the importance of the analysis unit selection in relating to particular learners' inflectional and derivational knowledge, the simple, lemma, and word-family counts' applicability has been explored. Based on Treffers-Daller et al.'s study, a lemma count seemed a more effective unit even for advanced L2 learners. If this is the case, a flemma count, which requires learners' higher inflectional knowledge (i.e., the inflectional knowledge regardless of the word class), might be a more discriminating unit for such high proficiency learners. Thus, I believe that a flemma count, which lies somewhere between lemma and word-family counts, might better capture their proficiency differences. However, no single study to date has examined the flemma count's usability in LD assessment.

Second, relating to the second factor of the L1 background influence on LD measures, several reviewed studies (Gonzalez, 2017; Read & Nation, 2006; Treffers-Daller et al. 2018; Zenker & Kyle, 2021) analyzed L2 English learners who represented a wide variety of nationalities. Nasseri and Thompson's (2021) study

investigated participants from Iran, but the authors mentioned that their first languages were different. Yu (2010) validated LD measures with L2 learners from 38 mixed L1 backgrounds. In doing so, he illuminated an important issue that needs more attention, i.e., L1 background influence on LD measures and scores, by examining the mixed L1 group as well as specific L1 groups (Filipino and Chinese). Three studies (Jarvis, 2013; Jarvis, 2017, McCarthy & Jarvis, 2013) controlled for L1 background, comparing L1 Finnish and L1 Swedish learners of English. Additionally, Zhang and Daller (2019) validated LD measures' applicability in the Chinese context. However, further investigations and validations of the L1 background effects on LD scores and measures are needed to gain more information on the word unit selection for specific L1 backgrounds.

The third factor relates to the L2 language proficiency influence on LD measures. LD measures' predictive power of different proficiency levels (overall, writing, or speaking) have been sufficiently examined. Most reviewed studies in section 2.3 highlighted the extent to which LD measures were useful in predicting learner groups at different proficiency levels. However, we know little about LD measure predictions of within-group differences (i.e., within a single proficiency level) despite some experimental evidence of within-group lexical variations (Read & Nation, 2006).

To my knowledge, no study to date has yet provided any information on how and which LD measures could estimate the differences among learners within the same proficiency level. Intra-group differentiation seems important for those cases where L2 learners' proficiencies are not different enough to be classified as separate levels, so they fall within a single level. For instance, in my experience as a language teacher in a country where English is a foreign language, one of my writing classes

comprised mostly intermediate level writers and very few advanced writers. In that case, I need LD measures which are predictive of intra-group variations.

The fourth factor, text length, has long been reported as the important issue in LD assessment. Despite the emergence of more robust, sophisticated LD measures which show more stability (e.g., *D, MTLD, HD-D*), these measures remain affected by text length. However, most LD studies in section 2.3 examined written or spoken samples with differing lengths, with the sole exception of Treffers-Daller et al. (2018), who set the constant text length (200 words) in assessing the written LD role in predicting L2 general language proficiency. As an exception, McCarthy and Jarvis (2013) used a corpus that varied in word counts since the study aimed at investigating LD measures' ecological validity. Due to the LD measures' text sample size problem, future research should analyze LD scores of the same length texts for more reliable LD score comparisons.

Overall, the LD studies reviewed in section 2.3 together evidenced that LD measures can be useful language predictors when controlling one or two of these four important factors. To attempt to build on their findings and advance awareness and understanding of LD, in this dissertation, I explore how and how much LD measures rely on the analysis unit, L1 background, language proficiency, and text length, grounding the study on Treffers-Daller et al. (2018). These four factors are addressed in four experimental chapters as follows. The experimental chapters (3, 4, and 5) will examine LD measure predictions of IELTS-based writing proficiency, and the experiment reported in chapter 6 will explore the extent to which LD measures predict IELTS-based speaking proficiency.

Chapter 3 explores the effects of two factors (analysis unit and text length) on LD measures. I analyze written texts of the same length (200 words) produced by L2

English learners from mixed L1 backgrounds (N = 194) and investigate LD measure predictions of their different writing proficiency levels (IELTS 6.5, 7, 7.5) based on simple, flemma, and lemma counts. I attempt to answer two research questions in this chapter: (i) How do flemmatization and lemmatization influence LD scores and LD measures' discrimination between IELTS-based writing proficiency levels?; and (ii) To what extent do LD measures predict IELTS-based writing proficiency levels based on simple, flemma, and lemma counts?

Chapter 4 controls for three factors (analysis unit, L1 background, and text length). This chapter investigates LD measures' ability to predict the writing proficiency of L1 Chinese L2 English learners (N = 105), which was the majority L2 group separated from the entire population (N = 194). Text length was also controlled, using the same constant text length (200 tokens) as Treffers-Daller et al. (2018). The two research questions in this chapter are: (i) How do flemmatization and lemmatization influence LD scores and LD measures' discrimination between IELTS-based writing proficiency levels of L1 Chinese L2 English learners?; and (ii) To what extent do LD measures predict IELTS-based writing proficiency of L1 Chinese L2 English learners based on simple, flemma, and lemma counts?

Chapter 5 deals with all four factors: analysis unit, L1 background, text length, and language proficiency. I explore the variation in LD measure predictions of writing proficiency based on L1 Chinese L2 English learners' writing proficiency with the intention of providing information on what the potentially effective LD measures might be for each specific writing level (e.g., 6.5, 7, or 7.5). I examine the LD measure predictions of intra-group writing proficiency depending on the different analysis units while setting a consistent text length. The research questions for this chapter are: (i) How large is the writing variability within the three writing

proficiency levels (IELTS 6.5, 7, 7.5)?; and (ii) To what extent do LD measures predict writing proficiency of L1 Chinese L2 English learners for simple, flemma, and lemma counts?

Chapter 6, which examines LD measures' ability to predict L2 speaking proficiency, addresses only two factors (the analysis unit and text length influences) since the participant sample size (N = 55) was too small to limit L1 background and language proficiency. However, the initial data analysis results indicated that the extent to which LD measures predict speaking proficiency is not significant at 200-word constant text length as used in the writing experiments. This led me to conduct further analyses to explore the minimum constant text length at which LD measures showed stronger speaking proficiency predictions. This chapter answered two research questions: (i) Based on different analysis units and text lengths, how do flemmatization and lemmatization influence LD scores and LD measures' discrimination between speaking proficiency levels?; and (ii) Based on different analysis units and text lengths, to what extent do LD measures predict IELTS-based speaking proficiency levels?

2.6 Conclusion

This literature review has examined LD studies which have explored and established how and the extent to which LD measures can be useful in predicting language proficiency (general, writing, and speaking) as well as the studies' attempts to consider and control the four important influencing factors on LD measure predictions. The review highlights the need for a deeper examination of the potential utility of the flemma count to contribute to the ongoing debate on the optimum analysis unit selection in the L2 context. The review has highlighted that there is a

pressing need for LD research to incorporate and control these four factors. I attempt to respond to these important research gaps in the ensuing four experimental chapters.

Chapter 3

Investigating Different Analysis Units (Simple, Flemma, and Lemma Counts) Influences on LD Measure Predictions of L2 Writing Proficiency

3.1 Introduction

The literature review in chapter 2 highlighted a crucial potential factor influencing the extent to which LD measures' predictive power that has not been sufficiently addressed in LD assessment. The factor is how different word-counting criteria can variously influence LD measures. Even though LD measures have been shown to predict L2 language proficiency, researchers acknowledge that the word-counting criteria can influence LD measure predictions (Jarvis, 2017; Treffers-Daller, 2013; Treffers-Daller et al., 2018).

Recently, this issue of the analysis unit choice importance has gained greater prominence in LD assessment. Most LD studies discussed in 2.2 (Crossley & McNamara, 2013; Gonzalez, 2017; Nasseri & Thompson, 2021; Yu, 2009) focused on a simple count. With greater awareness of the importance of the analysis unit selection, Treffers-Daller et al. (2018) conducted a wide-ranging investigation into the analysis units. They indicated that different analysis units influence LD measure predictions. Treffers-Daller et al. investigated the LD measure predictions of highly proficient L2 learners' general proficiency, comparing three different analysis units (simple, lemma, and word-family counts). Their findings showed that these three analysis units each influenced LD measures in different ways. A lemma count could enable and enhance LD measures' predictive power of learners' general L2 language proficiency better than a simple or word-family count.

If a lemma count seemed a more impactful unit even for the advanced L2 learners in Treffers-Daller et al.'s study, then a flemma count, which lies somewhere

between lemma and word-family counts, might better capture proficiency differences. With this in mind, the current study hypothesized that highly proficient learners might be further able to distinguish between the word classes of inflections and thus investigates whether a flemma count might be a more impactful unit for optimizing LD measure predict such learners' proficiency.

As discussed in sections 2.2 and 2.3, despite individual papers examining the impact of the simple, lemma, and word-family counts in LD assessment, to my knowledge, no single empirical study has, as yet, reported on LD measure predictions of language proficiency based on a flemma count. Therefore, the current study represents the first study to investigate a flemma count's influence on LD measure predictions compared to simple and lemma counts.

3.2 Replicating Treffers-Daller et al. (2018)

The current study partially replicates Treffers-Daller et al. (2018), discussed in section 2.2. Treffers-Daller et al. established that we can use LD measures as a proxy for gauging general CEFR language proficiency; it also suggested a lemma count instead of simple or word-family counts as a more effective lexical unit for ascertaining the L2 learners' lexical diversity. Despite the merits of their study, it included some complicating factors that need further research.

First, Treffers-Daller et al. investigated how well the learners' written LD scores could predict their overall L2 language proficiency, encompassing all four language skills. Therefore, the current study questions how listening, reading, and speaking skill scores may have influenced their overall proficiency scores. The current study investigates the potential relationship between written lexical diversity and writing proficiency alone.

Second, Treffers-Daller et al. indicated that a lemma count, which requires only learners' inflectional knowledge under the same word class, was a more distinctive word-counting unit for the advanced L2 learners in their study than word-family count, which demands both inflectional and derivational knowledge. Treffers-Daller et al. examined different lexical units and recommended using a lemma count over simple and word-family counts. However, the authors did not explore a flemma count, which requires a slightly higher inflectional knowledge than a lemma count, as a potentially more discriminating counting method.

To address these two factors, the current replication study explored LD measures' predictive power of L2 writing proficiency based on a flemma count and compared to simple and lemma counts. The specific research questions for this chapter are:

RQ 1. How do flemmatization and lemmatization influence LD scores and LD measures' discrimination between IELTS-based writing proficiency levels?

RQ 2. To what extent do LD measures predict different IELTS-based writing proficiency levels based on simple, flemma, and lemma counts?

3.3 Current study

3.3.1 Participants

The participants in the study were 194 adult L2 learners of English enrolled on a pre-sessional academic English course at a UK university specializing in Humanities and Social Sciences. They provided access to their essays via the written consent form (see Appendix 1), and the Research Ethics Committee of Queen Mary University (UK) has approved this research (approval number: QMREC2414a). So that the essays were as comparable as possible, from the 554 students on the pre-

sessional course, a sub-group of students in the Humanities and Social Sciences pathway was selected. The participants were from 24 different L1 backgrounds (see Table 3.1), with most of the participants being either L1 Chinese (N=105), L1 Taiwanese (N=22), or L1 Thai (N=20). 95 participants (48.97%) were male and 99 (51.03%) were female. The classroom teachers considered the participants to be intermediate-level in their L2 writing. Their written language proficiency ranged from IELTS bands 6.5 (N=39), and 7 (N=81), to 7.5 (N=74). The IELTS bands represent the specific levels of competence used (see Appendix 2) by university departments to accept or reject international students. Table 3.2 shows the participants' writing proficiency according to their IELTS levels.

Table 3.1

Participants by L1 Backgrounds (N = 194)

Nationality	N	Nationality	N	Nationality	N
Chinese	105	Brazilian	3	Indian	1
Taiwanese	22	Mongolian	1	Greek	1
Thai	20	Saudi Arabian	1	Vietnamese	1
Turkish	9	Indonesian	1	Belgium	1
Japanese	6	Kuwaiti	1	Chilean	1
Italian	5	Jordanian	1	Hong Kong	1
Colombian	4	Hungarian	1		
South Korean	3	Portuguese	1		
French	3	Mexican	1		

Table 3.2

Participants' IELTS-Based Writing Proficiency Levels

IELTS writing level	6.5	7	7.5	Total
N	39	81	74	194

3.3.2 Data and scoring

Over a three-week intensive academic English course, participants completed essays on a common theme: globalization's impact on society. Participants wrote essays of 2000 words (+/- 10%), following specific guidelines: having a coherent argument, a clear structure, appropriate information, accurate vocabulary, grammar, and consistent academic style. The class teachers rated participants' essays using IELTS style writing rubrics which evaluated task fulfilment, organization, coherence, language (grammar, vocabulary, punctuation) and referencing skills, and then assigned the essays to IELTS writing bands 6.5, 7, and 7.5.

To ensure reliability, all teachers were required to attend standardization sessions to familiarize themselves with the marking system before the course and then a moderation session after they completed the classes. In addition, attributed scores were then second marked (and, with significant score variation third marked) by other teachers of the same course. As there were too few essays in the higher and lower bands to create a representative sample, essays from the three intermediate bands (6.5, 7, 7.5) were selected for analysis.

The content of the essays for the Humanities and Social Sciences students were all on the theme of globalization, within which there were four essay titles:

1. Critically assess the relationship between the processes associated with globalization and armed conflicts since the 1990s.

2. Ambitious, profit-minded global companies should still behave in a socially responsible way. Discuss.
3. Does Globalization lead to development? Discuss.
4. Can market forces and free competition alleviate global poverty? Discuss your answer with reference to corporate activity in poorer countries and contexts.

The essay titles were split equally among the students (25% to each title), with class time given (16 students per class) for working on essays under the supervision of their teacher. There was also an expectation that students would use out-of-class time (evenings and weekends) to research the literature and complete the writing.

In contrast to prior research on LD measures that had used short essays written under exam conditions with tight time restrictions, they wrote the essays in this study in conditions that more closely resembled the actual writing experience of undergraduate students at a UK university. While analyzing texts written in semi-controlled conditions is potentially problematic in that unintended variables may have crept in, we argue that this was more than compensated for by the authenticity of the writing experience that allows students to write to their full potential. As recent studies suggest (Csomay & Prades, 2018; Higginbotham & Reid, 2019), if we are to make valid generalizations about the vocabulary L2 learners use in their writing, then analysis needs to use texts that fully reflect the abilities of the participants.

3.3.3 Data processing

The data were cleaned, flemmatized, and lemmatized before analysis using the Python program (https://www.python.org), the Natural Language Toolkit (NLTK) Python package for natural language processing. First, in-text direct citations and direct quotations were excluded from the texts since the words used were directly copied from other sources and did not represent the participants' existing vocabulary

knowledge. Second, following Treffers-Daller et al.'s (2018) comparable procedure, proper names, acronyms, and cardinal numbers were removed to prevent LD score inflation. Moreover, the spelling errors found in the written texts were corrected, and the contracted forms were transformed into full forms (e.g., hasn't > has not, what's more > what is more).

3.3.4 Three different lemmatization techniques

I created three different text versions of each writing sample: non-lemmatized, flemmatized, and lemmatized texts. The non-lemmatized text versions were the original texts in which different words used were simply counted as different types. For flemmatized and lemmatized versions, the data were flemmatized and lemmatized using the Python program. Flemmatization reduced the inflected forms of the same words to the base words irrespective of what parts of speech they were. For example, the verb forms "*develop, develops, developed, developing*" and the adjectives "*developed*" and "*developing*" were converted into the base word "*develop*". Lemmatization converted the inflected forms of words under the same word class into the base words. For example, only the verb forms "*develop, develops, developed, developing*" were reduced to the base word "*develop*". Then, as in Treffers-Daller et al. (2018), the same consistent text length was set for the comparability of the findings. I took two hundred words from the middle of each essay in its three different text versions (non-lemmatized, flemmatized, and lemmatized) using the Gramulator (McCarthy et al., 2012). Table 3.3 shows an example text (25 tokens) of the three different text versions and the number of *Types* (different words) for each text version.

Table 3.3

Types and Token Scores of a Sample Text for Three (Non-Lemmatized, Flemmatized, and Lemmatized) Versions

Text version	Text	Tokens & Types
Non-lemmatized	the students are now playing tennis playing tennis is their favorite leisure activity and they always play tennis together some students are very good players	Tokens = 25 Types = 20
Flemmatized	the student be now play tennis play tennis be they favorite leisure activity and they always play tennis together some student be very good player	Tokens = 25 Types = 18
Lemmatized	the student be now play tennis playing tennis be they favorite leisure activity and they always play tennis together some student be very good player	Token = 25 Types = 19

3.3.5 Lexical diversity measures

To compute the LD scores for all three different text versions of each writing sample, the current study used the same LD measures as Treffers-Daller et al. (2018): three basic measures (*Types, TTR, Guiraud's Index*) and three sophisticated measures (*D, HD-D, MTLD*). *Types* and *TTR* values were calculated using the Lextutor Compleat Lexical Tutor (https://www.lextutor.ca). The *Guiraud's Index* was computed by dividing the number of *Types* by the square root of the number of *Tokens* (types/√tokens). *D* scores were calculated using D_Tool (Meara & Miralpeix, 2016) (http://www.lognostics.co.uk/tools/D_Tools/D_Tools.htm), and *MTLD* and *HD-D* scores were computed using the Tool for the Automatic analysis of Lexical Diversity **(**TAALED; Kyle, Crossley, & Jarvis, 2021) https://kristopherkyle.github.io/professional-webpage/docs/tools.html.

3.3.6 Statistical analyses

A Shapiro-Wilk test was applied to check whether the data (writing scores and six LD scores under the three different word counting criteria for three writing levels, i.e., IELTS 6.5, 7 and 7.5) met the normality assumption. The findings indicated that some of the data were skewed. A Box Plot test showed the dataset included 13 outliers (11 mild outliers and two extreme outliers). I included the mild outliers because their teachers still considered the score differences representative. The effects of the two extreme outliers (one student from the IELTS 7 group for non-lemmatized *HD-D* score and one student from the IELTS 7.5 group for writing score) on the findings and interpretations of the results were examined. Since the analyses with and without the extreme outliers revealed different findings (differences in *F* values and in the extent to which the LD measures predicted, particularly, *HD-D*), I excluded the extreme outliers from the analysis.

I explored the flemmatization and lemmatization effects on LD scores and LD measures' discrimination between writing levels by using two-way and one-way ANOVA analyses with Bonferroni alpha adjustment. I also conducted Pearson correlational analyses to examine the correlations between LD measures and writing scores. Furthermore, I investigated the extent to which LD measures predicted writing proficiency by using regression analyses, in which the LD measures that were most strongly correlated with writing scores served as the predictor variables. Additionally, I ran the post hoc power calculations for one-way ANOVA and regression analyses by using G* Power software (ver. 3.1.9.7).

3.4 Results

The current chapter examined the extent to which LD measures predict L2 writing proficiency was influenced by the use of different word-counting units (simple, flemma and lemma counts). First, the descriptive statistics of the LD scores were calculated on the three different text versions (non-lemmatized, flemmatized, lemmatized) for both basic and sophisticated measures, reported in Tables 3.4 and 3.5. Table 3.4 presents the LD scores calculated with basic LD measures (*Type, TTR, Guiraud's Index*), and Table 3.5 shows the LD scores computed with sophisticated measures (*D, MTLD, HD-D*). The observed G power values gained from the post hoc power analyses were also reported in Tables 3.4 and 3.5.

Table 3.4

Descriptive Statistics of Basic LD Measures

Measure	6.5	7	7.5	Overall	Eta squared	Observed Power
Types0	119.77 (9.24)	118.01 (8.68)	114.33 (9.60)	116.97 (9.36)	.054	.84
Types1	114.64 (7.51)	110.75 (8.04)	106.95 (8.84)	110.09 (8.70)	.108	.99
Types2	116.74 (7.33)	112.28 (8.20)	108.47 (9.26)	111.73 (8.95)	.116	1.00
TTR0	.60 (.05)	.59 (.04)	.57 (.05)	.59 (.05)	.052	.97
TTR1	.57 (.04)	.56 (.04)	.54 (.04)	.55 (.04)	.101	.97
TTR2	.59 (.04)	.56 (.04)	.54 (.05)	.56 (.04)	.119	1.00
Guiraud0	8.47 (.65)	8.35 (.61)	8.09 (.69)	8.27 (.67)	.051	.66
Guiraud1	8.11 (.53)	7.83 (.57)	7.56 (.63)	7.78 (.62)	.108	.99
Guiraud2	8.27 (.51)	7.94 (.58)	7.67 (.65)	7.90 (.63)	.122	1.00

Note: 0 = Simple count, 1 = Flemma count, 2 = Lemma count

Table 3.5

Descriptive Statistics of Sophisticated LD Measures

Measure	6.5	7	7.5	Overall	Eta squared	Observed Power
D0	82.32 (20.24)	77.42 (18.69)	69.05 (20.16)	75.23 (20.15)	.066	.92
D1	68.03 (13.74)	65.18 (15.05)	58.27 (14.28)	63.13 (14.96)	.070	.93
D2	72.73 (16.83)	67.03 (15.51)	60.81 (15.53)	65.82 (16.33)	.075	.95
MTLD0	69.32 (17.51)	66.61 (17.12)	60.92 (19.94)	65.00 (18.53)	.033	.61
MTLD1	67.17 (16.93)	65.57 (16.12)	59.03 (18.72)	63.40 (17.57)	.039	.70
MTLD2	70.22 (16.69)	66.84 (17.67)	60.11 (17.62)	64.97 (17.83)	.051	.82
HD-D0	.80 (.03)	.80 (.03)	.78 (.03)	.79 (.03)	.084	.99
HD-D1	.80 (.03)	.79 (.03)	.78 (.03)	.79 (.03)	.095	.87
HD-D2	.80 (.03)	.79 (.03)	.78 (.03)	.79 (.03)	.075	.87

3.4.1 Flemmatization and lemmatization influences on LD scores, and LD measures' discrimination between IELTS-based writing proficiency levels.

As Tables 3.4 and 3.5 show, flemmatization and lemmatization did affect LD scores and measures. For all three basic measures and two of the sophisticated measures (*D, MTLD*), the overall LD mean scores from non-lemmatized data in which a simple count was used to calculate different words, were the highest, followed by the LD mean scores from lemmatized data, and then the LD mean scores for flemmatized data. However, for the sole sophisticated measure of *HD-D*, the overall LD mean scores for the three lemmatization methods (non-lemmatization, flemmatization, and lemmatization) were almost the same.

Two-way ANOVA analysis revealed that the three different word-counting units differed statistically and significantly for all LD measures: *Types (F (2,378)=* 202.372, p=<.001), *TTR* (*F (2,378)=* 204.367, *p*=<.001), *Guiraud's Index* (*F (2,378)=* 208.768, *p*=<.001), *D* (F *(2,378)=* 199.890, *p*=<.001), *MTLD* (*F (2,378)=* 7.318,

p=<.001), and *HD-D* (*F (2,378)*= 7.855, *p*=<.001). For all three basic measures and *D*, the differences between simple, flemma, and lemma counts were significant; however, for *MTLD* and *HD-D*, the simple and lemma counts were not significantly different (see Table 3.6).

As shown by the Eta squared values in Tables 3.4 and 3.5, LD measures' ability to discriminate between different writing levels was influenced by the analysis unit choice. The higher Eta squared values indicated that flemmatization and lemmatization could enable and enhance LD measures' discrimination between writing levels better than the non-lemmatization technique. Conversely, though, *HD-D* alone was less discriminative of writing levels on lemmatized text than on non-lemmatized text.

With the highest Eta square values among the three different word-counting criteria, the lemma count had the strongest influence on LD measures' discriminative power of L2 writing levels. All three basic measures (*Types, TTR, Guiraud's Index*) and two of the sophisticated measures (*D, MTLD*) were better discriminators of writing levels based on a lemma count. In contrast, *HD-D* was a better writing indicator once a flemma count was applied. The finding that a lemma count was found to have the most significant effects on LD measures' discrimination between writing levels was consistent with Treffers-Daller et al.'s (2018) finding on using a lemma count being more appropriate than simple or word-family counts.

Treffers-Daller et al. (2018) explored the extent to which LD measures can predict general CEFR language proficiency based on the lemma count, which could better discriminate between CEFR scores than simple and word family counts. However, the current study explored the extent to which LD measures can predict L2 writing proficiency, specifically based on all three analysis units (simple, flemma, and

lemma counts) to glean more insights into the different analysis unit influences on whether LD measures can predict L2 writing proficiency.

Table 3.6

ANOVA and Post Hoc Test Results of the Overall Differences Between Simple, Flemma, and Lemma Counts

Measure	*F*	*p*	0 vs 1	0 vs 2	1 vs 2
Types	202.372	<.001	*	*	*
TTR	204.367	<.001	*	*	*
Guiraud	208.768	<.001	*	*	*
D	199.890	<.001	*	*	*
MTLD	7.318	<.001	*	NS	*
HD-D	7.855	<.001	*	NS	*

The analysis findings for all three analysis units indicated that basic measures (*Types, TTR, Guiraud's Index*) were only predictive of 6.5 and 7 levels when a lemma count was used. Furthermore, *MTLD* was only effective in predicting writing levels 6.5 and 7.5 when based on a lemma count.

When a simple count was applied, among all LD measures, *Types, TTR, D*, and *HD-D* could predict the highest (7.5) level and the two lower (6.5 and 7) levels (see Table 3.7) while *Guiraud's Index* could only be discriminative of 6.5 and 7.5 levels. However, none of the six LD measures could predict 6.5 and 7 levels, and *MTLD* could not predict any writing levels. The *F* values of *D* and *HD-D* were higher than that of all three basic measures, and, with the highest *F* value, *HD-D* was the best writing predictor.

As shown in Table 3.8, for a flemma count, both basic measures (*Types, TTR, Guiraud's Index*) and sophisticated measures (*D, HD-D*) could discriminate between the highest (7.5) and two lower (6.5 and 7) levels. However, *MTLD* could not predict any writing levels. Because of the higher *F* values, all three basic measures were stronger writing proficiency predictors than the sophisticated measures. With the highest *F* value, *Types* was the most discriminating LD measure among the three basic measures.

Considering all three word-counting criteria, the lemma count was the most influential analysis unit on LD measures (see Tables 3.7, 3.8, and 3.9). All three basic measures predicted all writing levels (6.5 vs 7 vs 7.5). As for sophisticated measures, *HD-D* could discriminate between 7.5 and two lower levels (6.5 and 7). However, *D* and *MTLD* were less predictive, being predictive of only the highest (7.5) and the lowest levels (6.5). With the higher *F* values, the basic measures were better L2 writing proficiency indicators, and, among them, *Guiraud's Index* was the most robust writing discriminator.

Table 3.7

ANOVA and Post Hoc Test Results for LD Measures (Simple Count) across Different Writing Levels

Measure	*F*	*p*	6.5-7	6.5-7.5	7-7.5
Types0	5.386	.005	NS	*	*
TTR0	5.200	.006	NS	*	*
Guiraud0	5.075	.007	NS	*	NS
D0	6.704	.002	NS	*	*
MTLD0	3.200	.043	NS	NS	NS
HD-D0	8.625	<.001	NS	*	*

Table 3.8

ANOVA and Post Hoc Test Results for LD Measures (Flemma Count) across Different Writing Levels

Measure	*F*	*p*	6.5-7	6.5-7.5	7-7.5
Types1	11.476	<.001	NS	*	*
TTR1	10.590	<.001	NS	*	*
Guiraud1	11.467	<.001	NS	*	*
D1	7.124	.001	NS	*	*
MTLD1	3.879	.022	NS	NS	NS
HD-D1	9.875	<.001	NS	*	*

Table 3.9

ANOVA and Post Hoc Test Results for LD Measures (Lemma Count) across Different Writing Levels

Measure	*F*	*p*	6.5-7	6.5-7.5	7-7.5
Types2	12.448	<.001	*	*	*
TTR2	12.713	<.001	*	*	*
Guiraud2	13.179	<.001	*	*	*
D2	7.632	<.001	NS	*	NS
MTLD2	5.052	.007	NS	*	NS
HD-D2	7.670	<.001	NS	*	*

Overall, Table 3.10 shows that the extent to which LD measures predict L2 writing proficiency depended on the analysis units. The basic measures were better writing indicators than the sophisticated measures, and a lemma count could best enable and enhance the basic measures' power to predict writing proficiency compared to simple and flemma counts. Among the sophisticated measures, *HD-D* could better discriminate between writing levels than *D*, whereas *MTLD* was the least predictive of writing proficiency, only being effective when based on a lemma count. The sophisticated measures became more discriminative of writing proficiency once simple and flemma counts were employed rather than a lemma count.

Table 3.10

Each LD Measure across Three Different Analysis Units

	Simple count			Flemma count			Lemma count		
	6.5-7	6.5-7.5	7-7.5	6.5-7	6.5-7.5	7-7.5	6.5-7	6.5-7.5	7-7.5
Types	NS	*	*	NS	*	*	*	*	*
TTR	NS	*	*	NS	*	*	*	*	*
Guiraud	NS	*	NS	NS	*	*	*	*	*
D	NS	*	*	NS	*	*	NS	*	NS
MTLD	NS	NS	NS	NS	NS	NS	NS	*	NS
HD-D	NS	*	*	NS	*	*	NS	*	*

3.4.2 Exploring the extent to which LD measures predict IELTS-based writing proficiency based simple, flemma, and lemma counts.

First, I conducted Pearson analyses to investigate the correlations between participants' LD scores and writing scores for all three word-counting units (simple, flemma, and lemma counts). The findings are reported in Tables 3.11, 3.12 and 3.13. For all three analysis units, LD measures were strongly and positively correlated, particularly the basic measures (r = .989 to 1.000). This finding indicates that LD measures can well measure the same construct of lexical diversity. On the other hand, LD measures showed low to moderate negative correlations with writing scores, implying that it cannot always be assumed that the higher the writing proficiency level, the greater the lexical diversity knowledge. For instance, as shown in Figure 3.1, the highest writing proficiency (7.5) group used less diverse vocabulary than the two lower (6.5 and 7) levels.

Once a simple count was applied, *Types* and *HD-D* were the LD measures most highly and negatively correlated with writing scores. For a flemma count, *Types, Guiraud's Index* and *HD-D* had the strongest negative correlations with writing. For a lemma count, *Guiraud's Index* and *D* were the most strongly and negatively correlated with writing.

Figure 3.1

Types-Token-Ratio Scores Across Three Different IELTS Writing Levels

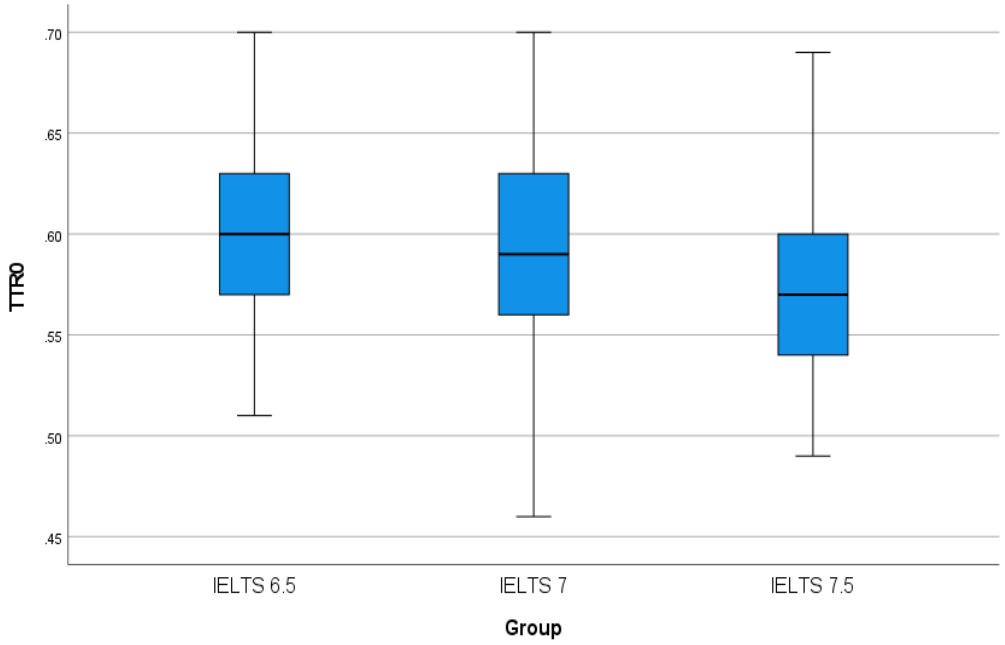


Table 3.11

Correlations between LD Scores and Writing (Simple Count)

	TTR0	Guiraud0	D0	MTLD0	HD-D0	Writing
Type0	.998**	.999**	.782**	.770**	.797**	-.243**
TTR0		.998**	.783**	.767**	.798**	-.238**
Guiraud0			.776**	.765**	.793**	-.237**
D0				.826**	.890**	-.263**
MTLD0					.876**	-.171*
HD-D0						-.272**

Table 3.12

Correlations between LD Scores and Writing (Flemma Count)

	TTR1	Guiraud1	D1	MTLD1	HD-D1	Writing
Type1	.989**	1.000**	.843**	.801**	.818**	-.314**
TTR1		.989**	.839**	.797**	.813**	-.306**
Guiraud1			.843**	.801**	.818**	-.314**
D1				.886**	.966**	-.251**
MTLD1					.863**	-.186**
HD-D1						-.289**

Table 3.13

Correlations between LD Scores and Writing (Lemma Count)

	TTR2	Guiraud2	D2	MTLD2	HD-D2	Writing2
Type2	.998**	.996**	.850**	.812**	.821**	-.341**
TTR2		.995**	.847**	.813**	.818**	-.344**
Guiraud2			.852**	.812**	.826**	-.348**
D2				.875**	.935**	-.280**
MTLD2					.860**	-.223**
HD-D2						-.279**

I further conducted regression analyses to investigate whether LD measures were valid writing proficiency predictors. However, it was deemed inappropriate to use all LD measures strongly correlated with each other as writing predictors due to the potential multicollinearity issues. Therefore, just the one basic and the one sophisticated measure that were each most strongly correlated with the writing scores were selected as the writing predictors, as Treffers-Daller et al. (2018) had done.

For the simple count, *Types* and *HD-D* served as the writing predictors. For the flemma count, among basic measures, *Types* and *Guiraud's Index* showed the same degree of correlation with writing. However, *Types* received the higher *F* value. Therefore, *Types* and *HD-D* were included in the regression analysis model as the writing predictors. *Guiraud's Index* and *D* were selected for the lemma count. Multicollinearity tests for all three analysis units showed no high correlations between the predictor variables in the same models because of the tolerance and VIF values falling within the limits.

Once a simple count was used, *Types* ($F(1,190) = 11.895$, $p = <.001$, $b = -.126$, $R^2 = .059$) and *HD-D* ($F(1,190) = 15.127$, $p = <.001$, $b = -41.331$, $R^2 = .074$) was

able to predict writing scores. *HD-D* (7.4%) could discern more variances in writing scores than *Types* (5.9%). However, these two measures turned insignificant when combined.

For a flemma count, both *Types* (F (1,190) = 20.796, p = <.001, b = -.176, R^2 = .099) and *HD-D* (F (1,190) = 17.329, p = <.001, b = -.45.401, R^2 = .084) were reliable predictors of writing scores, and *Types* (9.9%) was a better predictor than *HD-D* (8.4%). However, neither of these two measures could significantly predict writing scores when combined into the regression model.

With a lemma count, both *Guiraud's Index* (F (1,190) = 26.257, p = <.001, b = -2.675, R^2 = .121) and *D* (F (1,190) = 16.218, p =<.001, b = -.084, R^2 = .079) were significant and discerned 12.1% and 7.9% of the variances in writing scores respectively. When these two measures were combined together, *D* (p = .642) turned insignificant. When statistical power was analyzed using G power software, the power values were high for all these predictors across all three analysis units. Types and HD-D received high power values of .84 and .92 for no lemmatization, Types and HD-D received .98 and .95 for flemmatization and Guiraud's Index, and D received 1.00 and .94 values.

Overall, the findings of the regression analyses illustrated that a sophisticated measure, *HD-D*, was more powerful in predicting writing scores than a basic measure, *Types*, based on a simple count. However, *Types* was a more robust writing indicator than *HD-D* on the flemmatized data. *Guiraud's Index* could better discern the writing score variances than *D* on the lemmatized data. These findings indicate basic LD measures could be more effective in predicting writing scores once flemmatization and lemmatization were applied. In contrast, the sophisticated measure, *D*, seemed more predictive of writing proficiency when a simple count was used.

3.5 Discussion

The current study replicated Treffers-Daller et al. (2018) by investigating the extent to which different analysis units influence LD measure predictions of IELTS-based L2 writing proficiency. However, the current study differed from Treffers-Daller et al.'s study in two important ways. First, while they investigated the written LD role in predicting general proficiency covering four language skills, I based my investigation on the most related skill (writing), exploring the relationship between written LD and writing proficiency. Second, I examined a simple count and two alternative analysis units (lemma and flemma counts) that only require learners' inflectional knowledge.

The current study's findings support Treffers-Daller et al.'s (2018) assertion that lemmatization could best increase LD measure predictions of CEFR general proficiency. As expected, flemmatization and lemmatization influenced the extent to which LD scores and LD measures predicted writing proficiency. Due to the lemmatization process, LD scores on non-lemmatized data were the highest. After that, lemmatized data received higher LD scores than flemmatized data for most LD measures and for most writing levels.

Among the three analysis units (simple, flemma, and lemma counts), with the highest Eta squared values (see Tables 3.3 and 3.4), the lemma count had the strongest influence on LD measures, enhancing the predictive abilities of both basic and sophisticated measures, except the *MTLD*, which became a more distinguishing measure when based on a flemma count. A lemma count, requiring only learners' ability to consider the inflections of the same parts of speech, was a more distinctive analysis unit than a flemma count for the highly proficient L2 learners in the current study who might have inflectional knowledge irrespective of word classes.

The current study shows an important finding that the extent to which LD measures predict writing proficiency varies depending on the analysis units used. All three basic measures could best discriminate between writing levels on lemmatized data. *Types* and *TTR* were similarly helpful in predicting writing once simple and flemma counts were applied, whereas *Guiraud's Index* was less effective on non-lemmatized data when a simple count was used. Among the sophisticated measures, *HD-D* could predict writing proficiency, irrespective of the analysis units used. However, *MTLD* was only predictive of writing on lemmatized data, but *D* was less effective on lemmatized data.

Regression analyses indicated that the sophisticated measure, *HD-D*, was a stronger writing predictor than the basic measure, *Types,* when a simple count was used. For a flemma count, *Types* could discern a larger percentage of writing score variances than *HD-D*, whereas *Guiraud's Index* was a more predictive measure once a lemma count was used. These findings indicate that basic measures could be more predictive of writing proficiency once the data were flemmatized and lemmatized, whereas sophisticated measures, particularly, *D* and *HD-D,* could be more suitable for non-lemmatized data.

The current study thus validates LD measures as reliable IELTS-based writing proficiency indicators. It highlights the more significant effects of a lemma count on LD measures' predictive power than either simple or flemma counts. The current study has two important implications for researchers investigating lexical diversity and writing proficiency predictions by using different LD measures. First, the study highlights the flemmatization and lemmatization effects on LD scores and measures' writing level discrimination, as well as the greater influence of lemmatization on most LD measures' discriminative power of writing. Therefore, the current study supports

the importance of lemmatization techniques in LD assessment (Myint Maw et al., 2022; Treffers-Daller, 2013; Treffers-Daller et al., 2018).

Second, the study shows the differential effects of the three analysis units on LD measures' ability to predict writing proficiency. Two of the sophisticated measures (*D, HD-D*) were more predictive of writing proficiency than basic measures when a simple count was used, based on the higher *F* values and the regression analyses. However, once flemma and lemma counts are used, the basic LD measures were more likely to predict writing proficiency than the sophisticated measures, although *MTLD* could also predict writing once the data were lemmatized. These findings provide valuable insights into the importance of carefully selecting the most suitable analysis unit, depending on the LD measures used. For example, it might be better for researchers to use basic LD measures and *MTLD* when applying a flemma or lemma count. However, for using the other sophisticated measures (*D, HD-D*), a simple count might be a more discriminating analysis unit.

3.6 Limitations

The current study includes at least three limitations: text sample size, writing task and prompt, and participants' L1 background. First, the cut-off-point (200 words) is perhaps too small to adequately represent the complexity of ideas formulated in the lengthier essays used in the current study. I intentionally used the same consistent text length as Treffers-Daller et al. (2018) for comparability. However, it might be better for future studies to select longer texts, or systematically sample from the beginning, middle, and conclusion of the essays because of the potential differences in LD scores inherent in shorter and longer texts.

A second limitation relates to the essay writing task and prompt. The current study used a planned writing task, and this may have influenced participants' different

word use. For example, at the lowest level (6.5), participants might have chosen (by consulting dictionaries and thesauruses) advanced words beyond their existing vocabulary knowledge, resulting in similar degrees of lexical diversification to the higher-proficiency participants. Future studies should analyze the effect of spontaneous writing tasks on lexical diversity. The current study's participants attempted one of four essays on the same theme of globalization. Several earlier studies have shown that writing prompts can strongly influence LD scores (e.g., Alexopoulou et al., 2017; Kyle et al., 2016; Reid, 1986).Thus, future studies should control for a single essay prompt.

The third limitation relates to participants' diverse L1 backgrounds. Yu (2010) suggested that L1 background influences LD measures' ability to predict writing proficiency. The participants in the current study were from a wide variety of L1 backgrounds; therefore, there might have been the L1 background influences on the participants' lexical diversification, the LD measures' ability to predict writing, and the suitable analysis unit.

3.7 Conclusion

The current study investigated whether LD measures predict IELTS writing proficiencies under controlled text length conditions. The study also analyzed the extent to which LD measures can predict writing proficiency is influenced by three different analysis units (simple, flemma, and lemma counts). The findings indicate that LD measures can indeed be used as accurate writing proficiency predictors and that lemmatization has an enhancing effect on all LD measures except *MTLD*. Each LD measure's propensity to predict writing depends on the word-counting unit used. Basic LD measures seem more appropriate when flemmatization and lemmatization are applied, while sophisticated LD measures appear more distinguishing measures

for non-lemmatized data analysis, except for *MTLD,* which also works well for lemmatized data analysis. Overall, the current study confirms the greater predictiveness of the basic LD measures and the greater usefulness of a lemma count than either simple or flemma counts in L2 lexical diversity assessment. However, the current study's findings are based on L2 learners from various L1 backgrounds; the results are therefore only applicable to LD studies examining L2 learners with diverse L1 backgrounds.

Chapter 4

Investigating the extent to which L1 Background Influences LD Measure Predictions of L2 Writing Proficiency Based on Simple, Flemma, and Lemma Counts

4.1 Introduction

The study's findings of the experiment in chapter 3 supported the applicability of LD measures as reliable L2 English writing proficiency indicators. Also, they evidence showing that the extent to which LD measure predictions of writing proficiency is influenced by the lexical unit used, i.e., simple, flemma, and lemma counts. However, chapter three highlighted three limitations that might have affected the study's findings: the small text sample size, the writing task and prompt, and the participants' varied L1 backgrounds.

The study intentionally used the same constant text length (200 words) as Treffers-Daller et al. (2018) for the comparability of the findings. An examination of the effects of the writing task and the prompt was out of the study's reach since there were insufficient participant numbers for each essay. However, I believe that the participants' broad, diverse L1 backgrounds (N = 24) might have influenced the findings of the LD measure predictions of writing and the appropriateness of the three lexical units (simple, flemma, and lemma counts) under examination. Treffers-Daller et al. (2018) showed that LD measures could be reliable CEFR general proficiency predictors, proposing the lemma count as the most discriminating analysis unit. However, these findings were based on participants of multi-L1 backgrounds.

In LD assessment, only a few empirical studies (e.g., Yu, 2010), to date have examined the extent to which the L1 background influences LD measures predictions of writing proficiency. To address this crucial gap in LD research (L1 background

effects on LD measures) raised in chapter three, the experiment in chapter four controls for L1 background by considering a single L1 background group. The 194 participants were from 24 different L1 backgrounds, including China, Taiwan, Thailand, Japan, and Brazil; however, only the Chinese group comprised sufficient participant numbers (N=105), with the other L1 background groups each being small in comparison (N=22 or fewer). Therefore, the current chapter examines LD measures' ability to predict the writing proficiency of L1 Chinese L2 English learners. It explores whether a simple, flemma, or lemma count is the most appropriate analysis unit for L1 Chinese L2 English learners. The current chapter addresses the following specific research questions:

RQ 1. How do flemmatization and lemmatization influence LD scores and LD measures' discrimination between writing proficiency levels of L1 Chinese L2 English learners?

RQ 2. To what extent do LD measures predict writing proficiency of L1 Chinese L2 English learners based on simple, flemma, and lemma counts?

4.2 Study

4.2.1 Participants

The current study controlled for L1 background. The sub-group of the L1 Chinese L2 learners of English (N = 105) was selected from the entire population (N = 194) of 24 different nationalities (see Table 3.1 in Chapter three). The participants specialized in Humanities and Social Sciences, 55 of the participants (52.38%) were male and 50 (47.62%) were female. They consented to use their essays (see Appendix 1), and the Research Ethics Committee of Queen Mary University (UK) has approved this research (approval number: QMREC2414a). The class teacher rated the

participants' essays by using the IELTS-based writing rubrics (see Appendix 2). For the current study, I classified the participants into three different IELTS bands based on their writing scores: 6.5 (n=29), 7 (n=43), and 7.5 (n=33) bands (see Table 4.1). The sample texts of IELTS writing proficiency levels 6.5, 7, and 7.5 can be seen in Appendix 4.

Table 4.1

Chinese Participants' IELTS Writing Proficiency Levels

IELTS writing level	6.5	7	7.5	Total
N	29	43	33	105

4.2.2 Procedures

The current study adopted the same procedures and tools (Gramulator, Lextutor Compleat Lexical Tutor, D_Tool, TAALED) as for the entire data (N=194) analysis conducted in chapter three. The initial step, data treatment, included deleting the extraneous parts of each essay (in-text citations, direct quotations, tables, figures, punctation marks) and removing proper names, acronyms, and cardinal numbers to prevent LD score inflation. Spelling mistakes were also corrected.

As the second step, the data were flemmatized and lemmatized, again using the same Python package, and three different text versions (non-lemmatized, flemmatized, and lemmatized) were created for each writing sample. Then, the constant text length was set by again taking 200 words from the middle of each essay.

For the third step, each essay's LD scores for the three different text versions were computed using the same six LD measures: three basic measures (*Types, TTR, Guiraud's Index*) and three sophisticated measures (*D, MTLD, HD-D*).

4.2.3 Statistical analyses

I used a Shapiro-Wilk test to test the normality of the data (writing scores and six LD scores, each calculated based on three different word-counting criteria for the three writing levels, i.e., IELTS 6.5, 7 and 7.5). The findings showed that some data were not normally distributed. A Box Plot test showed that the skewed data had 15 outliers (two extreme outliers and 13 mild outliers). The writing 7.5 level group included one extreme outlier, and the *MTLD* scores for the 6.5 level included one extreme outlier. I included the mild outliers in the analysis since the class teachers still considered the score differences were acceptable. Data analysis both with and without the extreme outliers revealed different findings which affected the interpretation of LD measure predictions, particularly *D*. Because of these extreme outlier effects on the results, these two outliers were removed from the dataset.

To examine LD measures as writing proficiency predictors of L1 Chinese participants and to explore different analysis units' influences on LD scores and measures, the current study conducted the same statistical analyses used in the entire data (N=194) analysis for the experiment in chapter three. I analyzed LD scores by performing a series of Two-way and One-way ANOVA analyses with Bonferroni corrections, Pearson correlational analyses, and regression analyses. Furthermore, I conducted G power analyses for one-way ANOVA and regression tests.

4.3 Results

The current study investigated the extent to which LD measures predict IELTS-based writing proficiency of L1 Chinese participants depending on three different analysis units (simple, lemma and flemma counts). First, the descriptive statistics of the LD scores calculated on three different text versions (non-lemmatized,

flemmatized, lemmatized) were reported. Table 4.2 presents the LD scores calculated with basic LD measures (*Types, TTR, Guiraud's Index*), and Table 4.3 gives the LD scores computed with sophisticated measures (*D, MTLD, HD-D*). I also report the observed G statistical power values in Tables 4.2 and 4.3.

Table 4.2

Descriptive Statistics of Basic LD Measures

Measure	6.5	7	7.5	Overall	Eta squared	Observed Power
Types0	118.54 (7.88)	118.12 (8.03)	114.25 (10.04)	117.03 (8.78)	.046	.48
Types1	113.57(6.50)	110.98 (7.46)	106.75 (9.17)	110.37 (8.17)	.106	.87
Types2	115.29 (7.00)	111.86 (7.32)	108.34 (9.55)	111.70 (8.36)	.101	.85
TTR0	.60 (.04)	.59 (.04)	.57 (.05)	.59 (.04)	.044	.76
TTR1	.57 (.04)	.56 (.04)	.54 (.05)	.55 (.04)	.095	.76
TTR2	.58 (.04)	.56 (.04)	.54 (.05)	.56 (.04)	.104	1.00
Guiraud0	8.39 (.56)	8.35 (.57)	8.09 (.72)	8.28 (.62)	.043	.44
Guiraud1	8.03 (.46)	7.85 (.53)	7.55 (.65)	7.80 (.58)	.106	.87
Guiraud2	8.18 (.48)	7.90 (.52)	7.66 (.68)	7.90 (.59)	.111	.90

Table 4.3

Descriptive Statistics of Sophisticated LD Measures

Measure	6.5	7	7.5	Overall	Eta squared	Observed Power
D0	79.57 (16.41)	78.26 (17.53)	68.95 (21.38)	75.72 (18.92)	.059	.59
D1	66.04 (11.49)	66.37 (13.89)	58.14 (14.97)	63.72 (14.03)	.072	.69
D2	70.19 (16.33)	68.27 (13.67)	60.86 (16.33)	66.49 (15.61)	.062	.61
MTLD0	65.50 (13.10)	67.39 (16.54)	60.30 (20.65)	64.68 (17.23)	.031	.34
MTLD1	64.10 (12.37)	66.07 (14.44)	58.31 (19.59)	63.12 (15.94)	.044	.46
MTLD2	67.51 (13.95)	67.59 (16.17)	59.85 (18.55)	65.16 (16.63)	.046	.48
HD-D0	.80 (.02)	.80 (.03)	.78 (.03)	.79 (.03)	.089	.80
HD-D1	.80 (.02)	.80 (.03)	.78 (.03)	.79 (.03)	.110	.80
HD-D2	.80 (.03)	.80 (.03)	.78 (.03)	.79 (.03)	.065	.80

4.3.1 Flemmatization and lemmatization influences on LD scores and LD measures' discrimination between writing proficiency levels of L1 Chinese L2 English learners.

Flemmatization and lemmatization influenced LD scores and both basic and sophisticated LD measures. For both basic measures (Table 4.2) and sophisticated measures (Table 4.3), the non-lemmatized data where a simple count was used received the highest overall LD mean scores, followed by the lemmatized data, with flemmatized data receiving the lowest overall mean scores, except with *HD-D*. Moreover, the LD mean scores for each writing level showed the same pattern. For instance, *Types* scores on non-lemmatized data for writing 6.5, 7, and 7.5 levels were the highest, and *Types* scores on lemmatized data were higher than the scores on flemmatized data. The findings were the same for all of the other LD measures, except for *HD-D* scores, which were the same for all three text versions. These results indicate that the flemmatization and lemmatization processes that reduced the

inflections of words into the related base words influenced most LD measures and scores, with the sole exception of the *HD-D* scores.

Two-way analysis of variance revealed the statistically significant overall differences between the three different word-counting criteria for all LD measures: *Types (F (2,200)*= 119.714, p=<.001), *TTR* (*F (2,200)*= 123.767, *p*=<.001), *Guiraud's Index* (*F (2,200)*= 125.38, *p*=<.001), *D* (F *(2,200)*= 91.719, *p*=<.001), *MTLD* (*F (2,200)*= 4.468, *p*=.013), and *HD-D* (*F (2,200)*= 3.686, *p*=.027). For all basic measures and *D*, simple, flemma, and lemma counts differed significantly, whereas simple and lemma counts were not significantly different for *MTLD.* There were significant differences between only simple and flemma counts for *HD-D* (see Table 4.4).

The Eta squared values shown in Tables 4.2 and 4.3 indicate that these three different analysis units (simple, flemma, lemma counts) had different influences on the discriminative capacity of the different LD measures to gauge LD in writing. Because of the higher Eta squared values, flemmatization and lemmatization could better increase most LD measures' ability to discriminate between writing levels. Based on the Eta squared values, *Types*, *D*, and *HD-D* were better writing levels discriminators once the flemma count was applied, whereas *TTR*, *Guiraud's Index*, and *MTLD* could discern more writing variances based on the lemma count. Treffers-Daller et al. (2018) examined the extent to which LD measures predict CEFR overall language proficiency based on a single analysis unit, a lemma count, which could best enhance most LD measures predictions. However, the current study investigated three word-counting criteria in further analyses for more useful information on the appropriate selection of the analysis units depending on the LD measures used.

Table 4.4

ANOVA and Post Hoc Test Results of the Overall Differences between Simple, Flemma, and Lemma Counts

Measure	*F*	*p*	0 vs 1	0 vs 2	1 vs 2
Types	119.714	<.001	*	*	*
TTR	123.767	<.001	*	*	*
Guiraud	125.382	<.001	*	*	*
D	91.719	<.001	*	*	*
MTLD	4.468	.013	*	NS	*
HD-D	3.686	.027	*	NS	NS

Once we consider all three word-counting criteria, we observe that neither the basic nor the sophisticated measures could discriminate between the 6.5 and 7 IELTS writing levels, and *MTLD* was not predictive of writing proficiency at all, regardless of the analysis unit used. First, when a simple count was applied, among all LD measures, only *HD-D* could discriminate between the highest (7.5) level and the two slightly lower levels (6.5 and 7) (see Table 4.5).

Second, as shown in Table 4.6, once a flemma count was applied, all three basic measures were only effective in discriminating between the lowest and highest writing levels. Among the sophisticated measures, *D* could differentiate between the highest (7.5) and second highest (7) levels, whereas *HD-D* could predict the highest (7.5) and two lower (6.5 and 7) levels. With the highest *F* value, *HD-D* was the strongest writing indicator.

Third, based on the lemma count, only the basic measures (*Types, TTR, Guiraud's Index*) could discriminate between the 6.5 and 7.5 writing levels (see Table 4.7). In contrast, none of the three sophisticated measures were discriminative of any

writing levels. *Guiraud's Index* was the most powerful discriminator of writing groups with the highest *F* value.

Table 4.5

ANOVA and Post Hoc Test Results for LD Measures (Simple Count) across Different Writing Levels

Measure	*F*	*p*	6.5-7	6.5-7.5	7-7.5
Types0	2.408	.095	NS	NS	NS
TTR0	2.282	.107	NS	NS	NS
Guiraud0	2.242	.112	NS	NS	NS
D0	3.141	.048	NS	NS	NS
MTLD0	1.618	.203	NS	NS	NS
HD-D0	4.863	.010	NS	*	*

Table 4.6

ANOVA and Post Hoc Test Results for LD Measures (flemma count) across Different Writing Levels

Measure	*F*	*p*	6.5-7	6.5-7.5	7-7.5
Types1	5.931	.004	NS	*	NS
TTR1	5.238	.007	NS	*	NS
Guiraud1	5.924	.004	NS	*	NS
D1	3.894	.024	NS	NS	*
MTLD1	2.302	.105	NS	NS	NS
HD-D1	6.174	.003	NS	*	*

Table 4.7

ANOVA and Post Hoc Test Results for LD Measures (Lemma Count) across Different Writing Levels

Measure	*F*	*p*	6.5-7	6.5-7.5	7-7.5
Types2	5.635	.005	NS	*	NS
TTR2	5.787	.004	NS	*	NS
Guiraud2	6.236	.003	NS	*	NS
D2	3.283	.042	NS	NS	NS
MTLD2	2.437	.093	NS	NS	NS
HD-D2	3.452	.035	NS	NS	NS

Overall, the LD measures were influenced by the analysis unit employed, and different analysis units had different influences on LD measures' abilities to distinguish between writing proficiency levels (see Table 4.8). None of the three basic measures could discriminate between any writing levels when a simple count was applied; however, the basic measures were discriminative when flemma and lemma counts were applied. Among the sophisticated measures, *D* was only effective on flemmatized data, and *HD-D* could discriminate between writing levels when simple and flemma counts were used. However, *MTLD* could not predict writing through any of the three analysis units.

Table 4.8

Each LD Measure across Three Different Analysis Units

	Simple count			Flemma count			Lemma count		
	6.5-7	6.5-7.5	7-7.5	6.5-7	6.5-7.5	7-7.5	6.5-7	6.5-7.5	7-7.5
Types	NS	NS	NS	NS	*	NS	NS	*	NS
TTR	NS	NS	NS	NS	*	NS	NS	*	NS
Guiraud	NS	NS	NS	NS	*	NS	NS	*	NS
D	NS	NS	NS	NS	NS	*	NS	NS	NS
MTLD	NS	NS	NS	NS	NS	NS	NS	NS	NS
HD-D	NS	*	*	NS	*	*	NS	NS	NS

4.3.2 Exploring the extent to which LD measures predict writing proficiency of L1 Chinese L2 English learners based on simple, flemma, and lemma counts.

I performed Pearson analyses for all three analysis units to examine the correlations between LD measures and writing scores. As Tables 4.9, 4.10, and 4.11 show, for all three analysis units, the findings indicate that LD measures had strong and positive correlations with each other, particularly the basic measures, but low-to-moderate negative correlations with writing scores. The writing scores had the highest correlations with *Types* and *HD-D* for both simple and flemma counts and the highest correlations with *Guiraud's Index* and *D* for lemma count.

Table 4.9

Correlations between LD Scores and Writing (Simple Count)

	TTR0	Guiraud0	D0	MTLD0	HD-D0	Writing
Types0	.998**	.999**	.702**	.748**	.779**	-.198*
TTR0		.997**	.705**	.751**	.781**	-.191
Guiraud0			.696**	.744**	.774**	-.192
D0				.788**	.858**	-.220*
MTLD0					.872**	-.113
HD-D0						-.242*

Table 4.10

Correlations between LD Scores and Writing (Flemma Count)

	TTR1	Guiraud1	D1	MTLD1	HD-D1	Writing
Types1	.991**	1.000**	.796**	.763**	.768**	-.295**
TTR1		.991**	.791**	.761**	.761**	-.278**
Guiraud1			.797**	.762**	.768**	-.294**
D1				.878**	.971**	-.210*
MTLD1					.858**	-.131
HD-D1						-.268**

Table 4.11

Correlations between LD Scores and Writing (Lemma Count)

	TTR2	Guiraud2	D2	MTLD2	HD-D2	Writing
Types2	.997**	.992**	.809**	.778**	.781**	-.288**
TTR2		.990**	.805**	.779**	.778**	-.289**
Guiraud2			.813**	.778**	.791**	-.302**
D2				.860**	.933**	-.231*
MTLD2					.865**	-.166
HD-D2						-.222*

I then conducted regression analyses to examine whether LD measures were predictive of writing proficiency by using the one basic measure and the one sophisticated measure that most strongly correlated with the writing scores for each of the three analysis units. Collinearity tests indicated that multicollinearity was not a concern for all three analysis units since the tolerance and VIF values were within the limits.

For a simple count, the results indicated that *HD-D* (F (1,101) = 6.287, p = .014, b = -37.181, R^2 = .059) was significant and could discern 5.9% of the writing variances, whereas *Types* (F (1,101) = 4.127, p = .045, b = -.107, R^2 = .039) was not significant. However, when combined, these two measures were not significant.

For a flemma count, both *Types* and *HD-D* were predictive of writing scores, and *Types* (F (1,101) = 9.602, p = .003, b = -.171, R^2 = .087) could discern more writing score variances than *HD-D* (F (1,101) = 7.801, p = .006, b = -43.002, R^2 =.072). When these two measures were combined, neither was significant.

For a lemma count, both *Guiraud's Index* (F (1,101) = 10.153, p = .002, b = -2.415, R^2 = .091) and *D* (F (1,101) = 5.709, p = .019, b = -.070, R^2 = .053) could

significantly predict writing and discern 9.1% and 7% of the writing score variances. Again though, when these two measures were combined, *D* was not significant. When statistical power was calculated, G power analysis indicated that Types and HD-D received the values of .36 and .53 for simple count, Types and HD-D received .73 and .63 for a flemma count, and Guiraud's Index and D received -.76 and -.48 for a lemma count.

Overall, regression analyses indicated that *HD-D* was a more useful writing predictor than *Types* when a simple count was used. When a flemma count was used, both *Types* and *HD-D* were predictive of writing scores, with *Types* being the better writing predictor. Based on a lemma count, *Guiraud's Index* and *D* could significantly predict writing scores, with *Guiraud's Index* being better able to explain the writing score differences.

Comparing the relative LD measure predictions of L2 English learner with the mixed L1 background writing and that of a single L1 background group (Chinese) writing, as I expected, the extent to which LD measure predictions of writing is influenced by L1 background. Table 4.12 summarizes and compares the two experimental chapters' (3 and 4) findings on the extent to which LD measures predict writing based on simple, flemma, and lemma counts.

Table 4.12

Basic and Sophisticated LD Measure Predictions of Writing Proficiency (Mixed L1 & L1 Chinese)

		Simple count			Flemma count			Lemma count		
		6.5-7	6.5-7.5	7-7.5	6.5-7	6.5-7.5	7-7.5	6.5-7	6.5-7.5	7-7.5
Types	Mixed	NS	*	*	NS	*	*	*	*	*
	Chinese	NS	NS	NS	NS	*	NS	NS	*	NS
TTR	Mixed	NS	*	*	NS	*	*	*	*	*
	Chinese	NS	NS	NS	NS	*	NS	NS	*	NS
Guiraud	Mixed	NS	*	NS	NS	*	*	*	*	*
	Chinese	NS	*	NS	NS	*	NS	NS	*	NS
D	Mixed	NS	*	*	NS	*	*	NS	*	NS
	Chinese	NS	NS	NS	NS	NS	*	NS	NS	NS
MTLD	Mixed	NS	NS	NS	NS	NS	NS	NS	*	NS
	Chinese	NS	NS	NS	NS	NS	NS	NS	NS	NS
HD-D	Mixed	NS	*	*	NS	*	*	NS	*	*
	Chinese	NS	*	*	NS	*	*	NS	NS	NS

When L2 learners from various L1 backgrounds were examined, both basic and sophisticated LD measures were more discriminative of writing levels than were the same analyses for a single L1 Chinese group. For the multi-L1 group, *Types* and *TTR* could best discriminate between writing levels, whereas *HD-D* was discriminative of more writing levels for the L1 Chinese group. The more accurate predictive power of LD measures for the mixed L1 group might be due to the participants having various L1 backgrounds.

The findings of the most appropriate analysis units for these two analyses were different. For the multi-L1 background group analysis, the lemma count could best increase most LD measure predictions of writing proficiency. However, for the L1 Chinese group analysis, all LD measures except *MTLD* could discriminate more writing groups when based on the flemma count.

4.4 Discussion

To address an important limitation in chapter 3 (i.e., the potential L1 background influence on the extent to which LD measures predict writing), the current study restricted the L1 background to only L1 Chinese. It examined LD measures' capacity to predict the L1 Chinese participants' writing proficiency and examined the most discriminating analysis unit for these L1 Chinese L2 English writers. Similar to the multi-L1 background group analysis in chapter 3, the current study indicated similar results to Treffers-Daller et al. (2018). They found that the lemmatization technique influenced LD scores and measures and that LD measures were reliable CEFR general proficiency indicators.

First, the current study's findings indicate that flemmatization and lemmatization influenced LD scores and LD measures' discriminative power of writing levels. As expected, flemmatization and lemmatization processes lowered LD scores, leaving the highest LD scores on the non-lemmatized data, followed by the LD scores for the lemmatized data. And then, the LD scores for the flemmatized data were the lowest. The Eta squared values shown in Tables 4.2 and 4.3 highlight the more significant effects of flemma and lemma counts on the LD measures' ability to discriminate between writing levels. Based on the Eta squared values, *Types*, *D*, and *HD-D* were more powerful discriminating measures on flemmatized data, whereas

TTR, *Guiraud's Index,* and *MTLD* were stronger writing discriminators on lemmatized data.

Furthermore, the findings show that different word-counting criteria had different impacts on the extent to which LD measures predicted writing. When a simple count was used, only *HD-D* could discern the writing score differences. When a flemma count was used, all three basic measures (*Types, TTR, Guiraud's Index*), as well as two sophisticated measures (*D*, *HD-D*) were effective in estimating writing levels, and, with the highest *F* value, *Types* was the strongest writing indicator. Only the basic LD measures could differentiate between writing groups when a lemma count was used, with *Guiraud's Index* being the most powerful measure. *MTLD* failed to discriminate between any writing levels irrespective of the analysis unit used. Overall, flemmatization had a more significant influence on all LD measures than lemmatization, except for *MTLD*.

Regression analyses indicate that when a simple count was used, the sophisticated measure, *HD-D,* was predictive of writing proficiency, whereas the basic measure *Types* was not a useful writing predictor. Based on a flemma count, both *Types* and *HD-D* were predictive of writing scores; however, *Types* could discern more writing score variances. Both *Guiraud's Index* and *D* could predict writing scores based on a lemma count, with *Guiraud's Index* being a better writing predictor.

Stoeckel et al. (2020) reported that their low-level L1 Japanese L2 English learners with knowledge of just one word class could not comprehend other word classes, so they deemed a lemma count more representative. Similarly, Treffers-Daller et al.'s (2018) study indicated the lemma count's suitability for advanced L2 learners.

The mixed-L1 background group analysis (chapter 3) suggested using a lemma count over simple and flemma counts for L2 learners from intermediate to advanced levels.

The current study also supports the argument that LD measures can be used to predict IELTS-based writing proficiency. Regarding the analysis unit influence, flemma and lemma counts are found to have similar effects on the basic LD measures but different effects on the sophisticated measures, since only a flemma count could enhance the discriminating power of the sophisticated measures, except for *D*. The current study, therefore, suggests that a flemma count seems a more distinctive analysis unit than either a simple or lemma count in assessing the LD role in predicting writing proficiency of L1 Chinese L2 English learners.

Additionally, the current study supports Yu's (2010) findings on L1 background effects on whether LD measures predict writing proficiency. Comparing the results of chapters 3 and 4 (see Table 4.12), the current study provides insights into the varying L1 background influences on the predictive powers of LD measures and the selection of the most appropriate analysis unit for LD research in various L2 contexts. Most LD studies have examined L2 participant groups of various L1 backgrounds, possibly because of the limited data availability and/or the need for an adequate participant sample size for statistical power and validity. In such cases, though, researchers might need to select the suitable LD measure(s) and analysis units more carefully. For instance, *D* effectively predicts the writing proficiency of L2 learners from diverse L1 backgrounds regardless of the analysis unit used. However, researchers examining L1 Chinese participants should apply a flemma count if *D* is to be used.

4.5 Limitations

The current study has at least two limitations in need for further investigation: comparing different L1 background groups in LD assessment and exploring the language (L2 writing) proficiency effect on the extent to which LD measures predict writing proficiency and the analysis unit selection.

First, the current study attempts to control for L1 background and therefore examines L1 Chinese L2 English learners only. However, if the study could have compared different L1 background groups, it would have likely provided more detailed information on the influences of different L1 backgrounds (e.g., Chinese, Taiwanese, Thai) on the extent to which LD measures predict writing proficiency.

Second, despite the useful findings of the LD measure predictions of writing proficiency based on a single L1 background, the study has raised an important question that the existing studies in LD research have also not answered. The question relates to the participants' L2 writing proficiency influence on LD measures. There has been rich research on LD measures' ability to predict inter-group (i.e., different proficiency levels) variations in L2 writing ability, so this issue has been adequately addressed. However, to my knowledge, no single study has explored LD measures' intra-group differentiation to examine whether the same LD measure and the same analysis unit can be adopted for L2 English learners with the same L1 background but at different writing proficiency levels. This is the second limitation of this study, which could be a fecund area for future research.

4.6 Conclusion

The current study has examined the L1 background influence on LD measures' ability to predict the IELTS-based writing proficiency based on the simple,

flemma, and lemma counts when text length is controlled. The findings suggest that both flemma and lemma counts were better at enhancing most LD measures' capability (except *MTLD*) in estimating the writing proficiency of the L1 Chinese participants and that the LD measures can be useful writing predictors. Interestingly, LD measures' predictive abilities depended on the lexical unit used, indicating the greater suitability of the basic measures for both flemma- and lemma-based analyses but the sophisticated measures, particularly *HD-D*, for simple count analysis. However, the flemma count seemed a better analysis unit than simple or lemma count for L1 Chinese L2 English learners. In contrast, the lemma count was a more discriminating analysis unit for L2 learners with mixed L1 backgrounds, as shown in chapter 3 and in Treffers-Daller et al. (2018). The findings, therefore, support studies reporting L1 background influences on the extent to which LD measures predict writing proficiency and on the most suitable analysis unit choice for those particular L2 learners.

Chapter 5

Investigating Variation in the Extent to Which LD Measures Predict Writing Proficiency

5.1 Introduction

The experiment in chapter 4 validated the extent to which LD measures predict writing proficiency while controlling the three influential factors of word counting unit, L1 background, and text length. The study identified the most suitable analysis unit(s) and LD measures to predict the different writing proficiency levels (IELTS 6.5, 7, 7.5) of L1 Chinese L2 English writers. However, the study raised an important question which remains unanswered. If LD measures were discriminative of L1 Chinese participants at different writing proficiency levels, then LD measures could also be useful in discriminating between within-level writing proficiency differences, i.e., between the low and high sub-levels of each writing proficiency level. The sample written texts of low and high sub-levels can be seen in Appendix 5.

The current study, therefore, explores variation in the extent to which LD measures predict writing proficiency, and attempts to determine the most distinguishing analysis unit depending on the L1 Chinese participants' writing proficiency levels (IELTS 6.5, 7, or 7.5). Thus, the research questions for the current chapter are:

RQ 1. How large is the writing variability within the three writing proficiency levels (6.5, 7, 7.5)?

RQ 2. To what extent does variation in the LD measures' ability to predict writing proficiency depend on simple, flemma, and lemma counts?

5.2 Method

5.2.1 Participants

For the initial overall writing variability analysis, I investigated the same L1 Chinese L2 English writers (N = 103) at 6.5 (N = 28), 7 (N = 43), and 7.5 (N = 32), who were examined in chapter 4. They provided the written consensus to use their essays (see Appendix 1), and the Research Ethics Committee of Queen Mary University (UK) has approved this research (approval number: QMREC2414a). To analyze the extent to which LD measures predict intra-group writing variability, the participants in each writing (6.5, 7, 7.5) level were then sub-classified into two groups by using the K-means clustering method that allowed me to set the group numbers. Since the sample sizes of the three writing proficiency groups (6.5, 7, 7.5) were small, I decided to sub-classify each writing level into only two groups (low and high). These two clustered groups within each of the three writing levels were consistent, since ANOVA tests indicated that the groups differed significantly from each other within all three writing levels. Table 5.1 shows the participant number in each writing sub-level, and figures 5.1, 5.2, and 5.3 describe the clustered groups for the three writing proficiency levels.

Table 5.1

Writing Sub-levels of L1 Chinese Participants

IELTS Writing Level	Low	High	Total
6.5	9	19	28
7	24	19	43
7.5	16	16	32

Figure 5.1

Low and High Sub-Groups of Writing Proficiency 6.5 Level

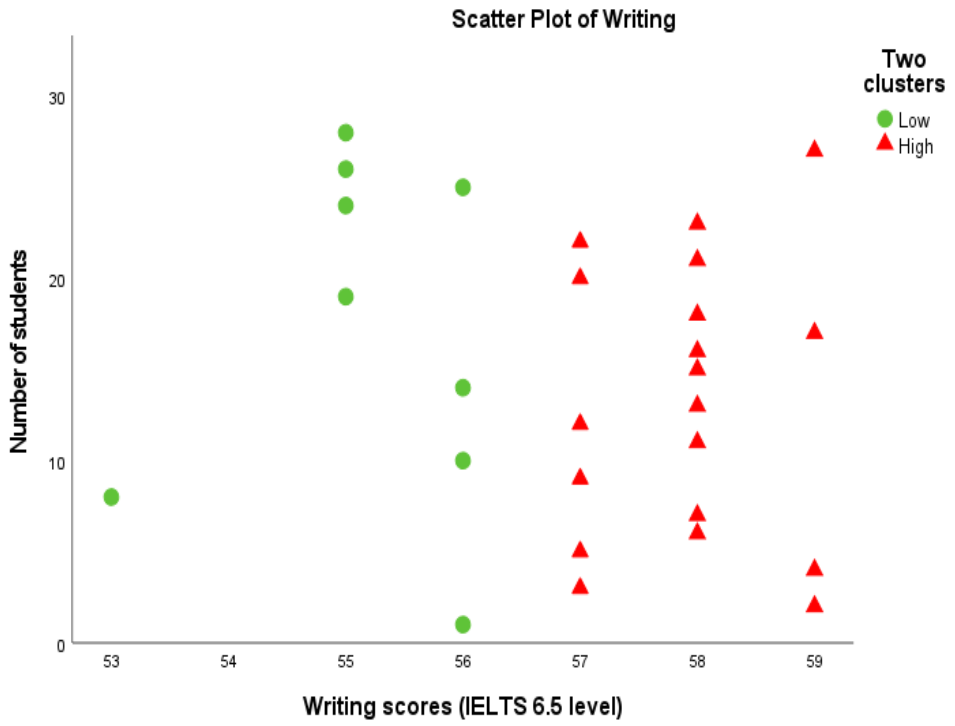


Figure 5.2

Low and High Sub-Groups of Writing Proficiency 7 Level

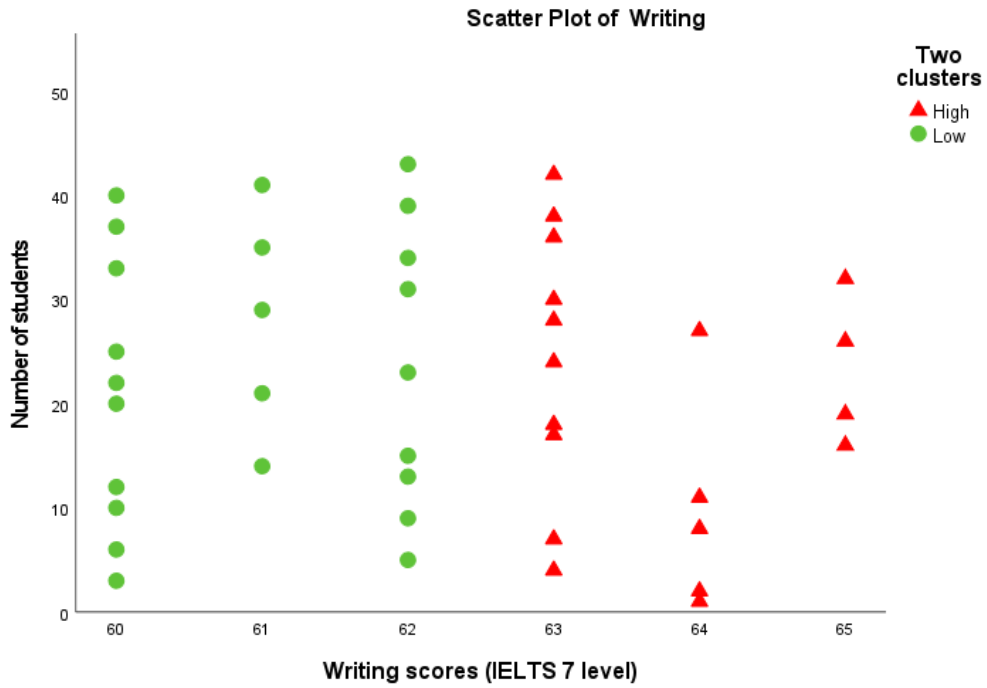
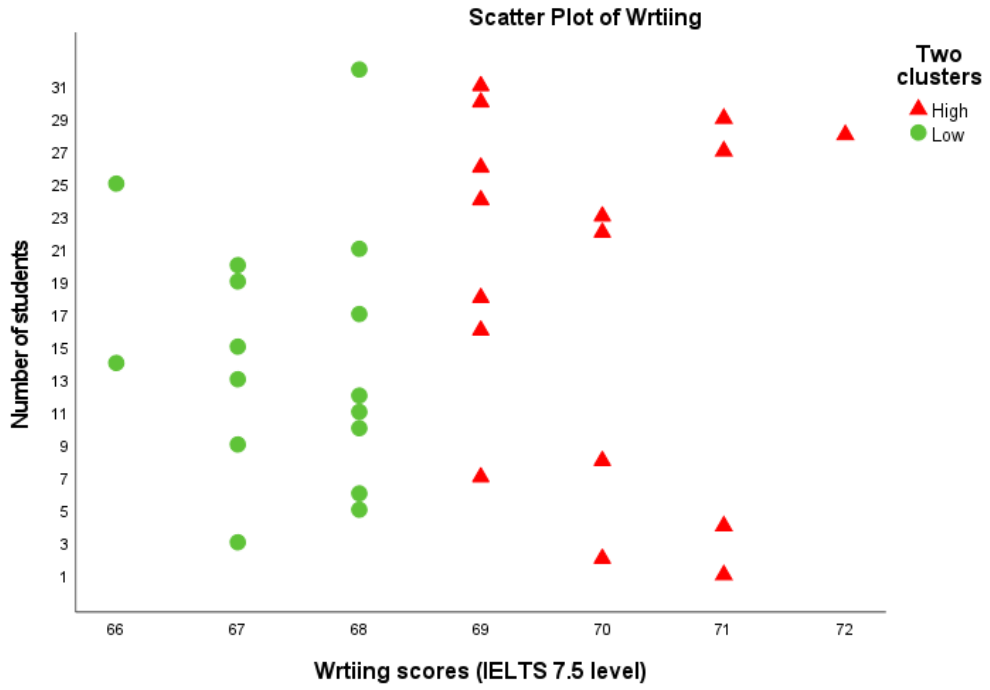


Figure 5.3

Low and High Sub-Groups of Writing Proficiency 7.5 Level



5.2.2 Statistical analyses

Before analysis, I examined the normal distribution of the data (IELTS-based writing scores and LD scores for all sub-groups across three analysis units) by using a Boxplot test. The findings indicated the data violated the normality assumption with both mild and extreme outliers. However, the participant sample sizes of low and high writing sub-groups were unbalanced and too small to remove the outliers, and also the class teacher considered these score differences still acceptable. I, therefore, kept all the outliers and applied non-parametric statistical tests for the skewed data throughout the study.

First, to explore the writing variability within the three different writing proficiency levels (6.5, 7, 7.5), I used both central tendency (Mean) and variability measures (Range, Interquartile range, Standard deviation, Mean absolute deviation,

and Variances) to accurately describe and summarize the data. The central tendency (Mean) indicated the score that most participant writers obtained; however, the mean score alone cannot indicate how the writers within a single writing level varied. Thus, to complement the central tendency measure, I also used the variability measures of Range (the difference between the lowest and highest scores), Interquartile range (IQR; the difference between the 3rd quartile and the 1st quartile), Standard deviation (SD; the average distance from the mean), Mean absolute deviation (MAD; the average distance between each writing score and the mean), and Variance (the average of the squared deviation from the mean). The inter-group variations of the three writing proficiency levels were compared using Fisher's two samples for variances (Fisher's F-test) to explore whether the variances were statistically and significantly different from each other. I then visually compared and summarized the writing variability by using comparative histograms.

Second, I explored LD measures' discrimination of writing variation within each sub-level (6.5, 7, 7.5) of writing proficiency based on the three analysis units used in the preceding experiments (simple, flemma, and lemma counts). First, I sub-divided the participants in each writing level into low and high writing sublevels by using K-means clustering. Second, the LD scores calculated on each of the non-lemmatized, flemmatized, and lemmatized text versions were analyzed and compared using Friedman's Two-way ANOVA and a Mann-Whitney U test. I investigated the relationships between LD measures and writing by using Spearman's correlation tests. To determine the extent to which LD measures predict writing sub-levels (low and high), regression analyses were conducted with the one basic and the one sophisticated LD measure which indicated the highest correlations with the writing scores for each of the three analysis units (simple, flemma, and lemma counts). Since

the data were non-normal, having outliers, I performed log-transformed regression analyses. I also calculated the statistical power for Mann-Whitney U test and regression analyses by using G power software.

5.3 Results

5.3.1 Writing variability within the three IELTS-based writing proficiency levels (6.5, 7, 7.5).

Table 5.2 presents the summary statistics of the writing variability within the three writing proficiency levels, calculated with the central tendency and variability measures. The mean scores for all three writing levels were consistent across the three proficiency levels since the mean scores increased with the higher writing levels. The three writing levels had each small and similar ranges, considered the highest and the lowest writing scores, and the interquartile range which determined the difference between the third and first quartiles of the distribution. The small Range and IQR scores implied that all three writing levels had low variability. The small standard and the mean absolute deviations of all three writing levels also highlighted that the writing score dispersions were not large, being clustered around the means.

I also used the variance measure that calculated the average distance of all values within a group from the mean. The small variance scores of all three writing levels confirmed the data were low-variant and writing proficiency level 7 seemed to have the greatest data dispersion. Fisher's two samples for variances suggested that the writing variances of the three proficiency (6.5, 7, 7.5) levels were not statistically different (see Table 5.3).

Overall, the variability analysis findings indicated that the data were not varied, and there were no significant differences in writing variability between the

three writing proficiency levels. A reason might be that the participant sample sizes were small, so the writing scores were not different. I visually described data dispersions of all three writing levels with the comparative histograms in Figure 5.4.

Table 5.2

Summary Statistics of Writing Variability Within the Three Writing Proficiency Levels (6.5, 7, 7.5)

		6.5	7	7.5
Central tendency	Mean	57.04	62.16	68.66
Variability	Range	6 (59 - 53)	5 (65 - 60)	6 (72 - 66)
	IQR	2 (58 - 56)	2 (63 - 61)	2.75 (70 - 67.25)
	SD	1.503	1.617	1.558
	MAD	1.181	1.344	1.281
	Variance	2.258	2.616	2.426

Note. IQR = interquartile range; SD = standard deviation; MAD = mean absolute deviation

Table 5.3

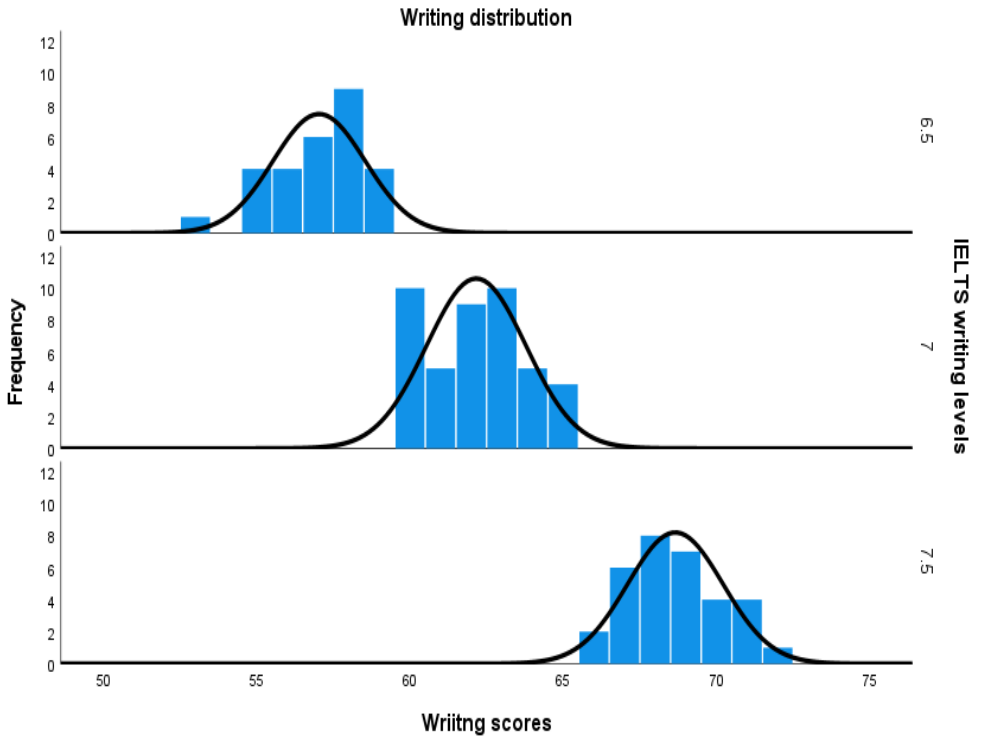
Comparison of the Writing Variances Between Three Proficiency Levels

	Fisher's F test		
	6.5 vs 7	6.5 vs 7.5	7 vs 7.5
F	.863	.931	1.078
F Critical	.547	.532	1.771
p	.348	.427	.419

Note. 6.5 (N = 28); 7 (N = 43); 7.5 (N = 32).

Figure 5.4

Visualization of Writing Variability Withing the Three Proficiency Levels (6.5, 7, 7.5)



5.3.2 Variation in the extent to which LD measures predict the writing proficiency of L1 Chinese L2 English learners.

To explore the extent to which variation in the LD measures' ability to predict writing proficiency depends on the word unit, I explored whether LD measures could discriminate between the writing proficiency sub-levels (low and high) based on the simple, flemma, and lemma counts. I analyzed three different writing proficiency levels (IELTS 6.5, 7, 7.5) separately, and I report the results for each level in the following three sub-sections (5.3.2.1 to 5.3.2.3).

5.3.2.1 Exploring the extent to which LD measures predict IELTS-based writing proficiency 6.5 sub-levels (low and high) of L1 Chinese L2 English learners based on simple, flemma, and lemma counts.

As for the descriptive statistics, I report the median scores because of the data skewness. Table 5.4 presents the median scores of the low (N = 9), and high (N = 19) IELTS 6.5 level sub-groups calculated on each of the non-lemmatized, flemmatized, and lemmatized text versions for both basic and sophisticated LD measures. The medians of the two sub-groups were similar for most LD measures.

Regarding the analysis unit's influence on the LD measures' intra-group writing discrimination, the Cohen's *r* values showed that the LD measures' discrimination was varied, depending on the word-counting criteria. Simple and flemma counts had greater effects on all three basic LD measures. As for the sophisticated measures, *D* and *MTLD* were more powerful on non-lemmatized and lemmatized versions, whereas *HD-D* was a more discriminative measure on lemmatized and flemmatized data. Interestingly, different analysis units again had different influences on the efficacy of the different LD measures. *Types* and *Guiraud's Index* were the most discriminating measures on the flemmatized data. *TTR* and *MTLD* were better writing discriminators based on the simple count, and *D* and *HD-D* were the most discriminative of writing sub-groups based on the lemma count.

Table 5.4

Medians and Cohen's r Values of Basic and Sophisticated LD Measures (IELTS 6.5 Level)

	Basic LD Measures					Sophisticated LD Measures			
	Low	High	*r*	Observed Power		Low	High	*r*	Observed Power
Types0	121	120	.079	.07	D0	73.30	77.70	.140	.09
Types1	114	114	.121	.30	D1	63.50	67.00	.056	.06
Types2	116	117	.028	.06	D2	66.00	68.30	.191	.45
TTR0	.61	.60	.080	.13	MTLD0	63.45	68.64	.051	.08
TTR1	.57	.57	.023	.13	MTLD1	61.13	64.08	.019	.05
TTR2	.58	.59	.014	.05	MTLD2	60.71	67.79	.042	.12
Guiraud0	8.56	8.49	.079	.07	HDD0	.80	.80	.038	.05
Guiraud1	8.06	8.06	.121	.29	HDD1	.80	.80	.094	.05
Guiraud2	8.20	8.27	.005	.05	HDD2	.80	.80	.173	.21

As shown in Table 5.5, first, the differences between the three analysis units for both basic and sophisticated LD measures were examined using a Friedman's Two-way ANOVA, a non-parametric test equivalent to a Two-way Repeated Measures ANOVA. The findings show that the simple, flemma, and lemma counts were statistically and significantly different from each other for all three basic measures. However, for the sophisticated measures, there were no significant differences between flemma, and lemma counts for *D*, and none of these three units differed significantly for *MTLD* and *HD-D*.

Next, to explore LD measures' discrimination between low and high writing sub-groups, I conducted Mann-Whitney U-tests for all six LD measures based on the three analysis units. The findings show that none of the LD measures could

discriminate between the IELTS 6.5 writing sub-groups, regardless of whether a simple, flemma, or lemma count was used (see Table 5.6).

Table 5.5

Friedman's Two-way ANOVA Results of the Overall Differences Between the Simple, Flemma, and Lemma Counts (IELTS 6.5 Level)

Measure	$\chi 2$ (2)	p	0 vs 1	0 vs 2	1 vs 2
Types	20.606	<.001	*	*	*
TTR	20.547	<.001	*	*	*
Guiraud	21.853	<.001	*	*	*
D	17.643	<.001	*	*	NS
MTLD	2.667	.264	NS	NS	NS
HD-D	6.169	.046	NS	NS	NS

Table 5.6

LD Measures' Discrimination Between the IELTS 6.5 Writing Sub-levels Based on the Three Analysis Units

Measure	Simple count			Flemma count			Lemma count		
	U	*p*	low - high	*U*	*p*	low - high	*U*	*p*	low - high
Types	77.00	.699	NS	72.50	.530	NS	82.50	.885	NS
TTR	77.00	.699	NS	83.00	.923	NS	84.00	.962	NS
Guiraud	77.00	.699	NS	72.50	.530	NS	85.00	1.000	NS
D	70.50	.468	NS	79.50	.772	NS	65.00	.332	NS
MTLD	80.00	.809	NS	83.50	.923	NS	81.00	.847	NS
HD-D	81.50	.847	NS	75.50	.629	NS	67.00	.383	NS

I further conducted Spearman correlation analyses to investigate the extent to which LD measures correlated with IELTS-based writing proficiency (see Table 5.7). The LD measures showed low to high positive correlations with each other, particularly the basic measures. However, as shown in Table 5.7, the correlations between the LD measures and writing scores were not consistent. For the simple count, all LD measures except *MTLD* had low positive correlations with writing scores. For both flemma and lemma counts, all three basic measures and *MTLD* had low positive correlations with writing, whereas *D* and *HD-D* showed low negative correlations with writing.

Among the basic measures, *Types* and *Guiraud's Index* were the most highly correlated with writing for the simple count, *TTR* had the highest correlations with writing for the flemma count, and *Types* was the most strongly correlated with writing for the lemma counts. As for the sophisticated measures, *D* and writing were the most

highly correlated for both the simple and lemma counts, but *MTLD* and writing indicated the highest correlation based on the flemma count.

Table 5.7

Correlations Between LD Measures and Writing (IELTS 6.5 Level)

		Types	*TTR*	*Guiraud*	*D*	*MTLD*	*HD-D*
Simple count	Writing	-.143	-.121	-.143	-.177	.159	-.071
Flemma count	Writing	.129	.135	.129	-.054	.179	-.106
Lemma count	Writing	.050	.040	.039	-.191	.035	-.179

I performed a series of regression analyses to examine the extent to which LD measures predict writing proficiency. Since the data were not normally distributed, I used the log-transformed regression analysis method. The writing was used as the outcome variable, and the one basic and the one sophisticated LD measure which had the strongest correlations with writing for each analysis unit were used as the predictor variables. The LD measures entered into the models met the multicollinearity assumption. The basic and sophisticated LD measures were first analyzed as separate writing predictors, and then measures were entered together into the model.

Based on the simple count, *Types* (F (1,26) = .095, p = .760, b = -.001, R^2 = .004) , *Guiraud's Index* (F (1,26) = .095, p = .760, b = -.011, R^2 = .004), and *D* (F (1,26) = .915, p = .348, b = -.001, R^2 = .034) could not predict writing when analyzed either separately or in combination. However, *D* seemed to be more predictive of writing than the basic measure, *Types*, since *D* could discern 3.4% of the writing score variances.

For the flemma count, neither *TTR* (*F* (1,26) = .154, *p* = .698, b = .068, R^2 = .006) nor *MTLD* (*F* (1,26) = .916, *p* = .347, b = .001, R^2 = .034) were statistically significant either as separate or as combined writing predictors. Similar to the simple count, it was the sophisticated measure, in this case *MTLD*, that could discern more writing variances (3.4%) than the basic measure, *TTR* (.06%).

For the lemma count, *Types* (*F* (1,26) = .081, *p* = .778, b = .001, R^2 = .003) and *D* (*F* (1,26) = 1.555, *p* = .224, b = -.001, R^2 = .056) could not significantly contribute to the model either when separated or combined. *D* was better able to predict writing (5.6%) than *Types* (.03%).

When the statistical power was calculated by using G power, the power values for *Types, Guiraud's Index* and *D* were .06, .06, and .16 based on simple count, .07 and .16 for *TTR* and *MTLD* based on flemma count, and .06 and .24 for *Types* and *D* based on lemma count respectively.

Overall, none of the LD measures analyzed were statistically significant either in separate or combined analyses. Interestingly, for all three analysis units, the sophisticated measures (*D, MTLD*) for each unit were all better able to estimate the writing variances because of their smaller *p* values and greater R^2 values than the basic measures (*Types, TTR*).

5.3.2.2 Exploring the extent to which LD measures predict IELTS-based writing proficiency 7 sub-levels (low and high) of L1 Chinese L2 English learners based on simple, flemma, and lemma counts.

Table 5.8 presents the median scores of the low and high sub-groups of IELTS writing proficiency 7 level and the observed statistical power values based on simple, flemma, and lemma counts. For most LD measures, the low group of the writing 7 level obtained higher medians than the high group. According to the Cohen's *r* values,

all three basic measures and *HD-D* were more effective in discriminating between low and high subgroups once non-lemmatization and lemmatization techniques were applied. *D* and *MTLD* were more powerful writing discriminators in the non-lemmatized and flemmatized text versions. Among all three analysis units, the simple count could best increase the predictions of writing proficiency of all LD measures except for *HD-D*, which was a better writing indicator based on the lemma count. Unlike for the IELTS 6.5 level, the findings of the three analysis units' influences were clear, showing that the simple count was the most impactful analysis unit on five out of the six LD measures under investigation.

The differences between simple, flemma, and lemma counts were examined using Friedman's Two-way ANOVA test (see Table 5.9). The findings indicate that the simple count differed significantly from flemma, and lemma counts for all three basic measures. The three analysis units differed significantly for *D* but not for *MTLD* and *HD-D*.

Table 5.8

Medians and Cohen's r Values of Basic and Sophisticated LD Measures (IELTS 7 Level)

Basic LD Measures					Sophisticated LD Measures				
	Low	High	*r*	Observed Power		Low	High	*r*	Observed Power
Types0	118.50	115.00	.086	.09	D0	82.15	70.00	.267	.47
Types1	112.50	108.00	.030	.05	D1	68.95	66.00	.205	.30
Types2	113.50	111.00	.054	.05	D2	71.60	67.10	.185	.20
TTR0	.60	.58	.082	.17	MTLD0	68.94	65.42	.086	.12
TTR1	.57	.54	.021	.05	MTLD1	67.17	63.52	.075	.07
TTR2	.57	.56	.047	.05	MTLD2	69.80	69.05	.007	.05
Guiraud0	8.38	8.13	.080	.10	HDD0	.81	.80	.204	.26
Guiraud1	7.96	7.64	.030	.06	HDD1	.81	.80	.175	.26
Guiraud2	8.03	7.85	.062	.06	HDD2	.81	.79	.207	.26

Table 5.9

Friedman's Two-way ANOVA Results of the Simple, Flemma, and Lemma Counts (IELTS 7 Level)

Measure	$\chi 2$ (2)	*p*	0 vs 1	0 vs 2	1 vs 2
Types	68.646	<.001	*	*	NS
TTR	70.235	<.001	*	*	NS
Guiraud	67.988	<.001	*	*	NS
D	59.209	<.001	*	*	*
MTLD	.235	.889	NS	NS	NS
HD-D	5.700	.058	NS	NS	NS

As shown in Table 5.10, to investigate LD measures' discrimination between writing sublevels, the low and high sub-groups of IELTS writing 7 were analyzed using the Mann-Whitney U test, which uses the ranks of the values instead of the mean scores for non-normally distributed data. The analyses showed that none of the six LD measures were discriminative of writing across all three analysis units. However, having looked at the *p* values, *D* and *HD-D* seemed more predictive of writing because of their smaller *p* values compared to the other LD measures.

Table 5.10

Mann-Whitney U-test Results of the LD Measure Discrimination Between IELTS 7 Writing Sub-levels

Measure	Simple count			Flemma count			Lemma count		
	U	*p*	low - high	*U*	*p*	low - high	*U*	*p*	low - high
Types	205.00	.573	NS	220.00	.845	NS	213.50	.722	NS
TTR	206.00	.588	NS	222.50	.892	NS	215.50	.759	NS
Guiraud	206.50	.598	NS	220.00	.845	NS	211.50	.686	NS
D	156.50	.080	NS	173.00	.179	NS	178.50	.226	NS
MTLD	205.00	.574	NS	208.00	.625	NS	226.00	.961	NS
HDD	174.00	.182	NS	181.50	.251	NS	173.00	.174	NS

I further analyzed LD measures and writing relationships using Spearman's correlation tests (see Table 5.11). There were low to strong positive correlations between LD measures. However, all three basic measures, *D*, and *HD-D* had low negative correlations with writing for all three differently lemmatized versions. Although *MTLD* also indicated a low negative correlation with writing based on the

simple and flemma counts, it showed a low positive correlation with writing based on the lemma count.

The one basic and the one sophisticated LD measure which were the most highly correlated with writing for each analysis unit were chosen to be used as the writing predictors in the further regression analyses to prevent any multicollinearity issues. Writing had the strongest correlations with *TTR* and *D* for both the simple and flemma counts and with *Guiraud's Index* and *HD-D* for the lemma count. Overall, LD measures could not significantly contribute to the regression models for all three analysis units, and the sophisticated measures (*D, HD-D*) appeared more discriminative of writing compared to the basic measures (*TTR, Guiraud's Index*) in this context.

Table 5.11

Correlations Between LD Measures and Writing (IELTS 7 Level)

		Types	TTR	Guiraud	D	MTLD	HDD
Simple	Writing	-.030	-.033	-.027	-.167	-.059	-.092
Flemma	Writing	-.009	-.012	-.009	-.083	-.036	-.060
Lemma	Writing	-.007	-.007	-.011	-.036	.032	-.051

A series of log-transformed regression analyses, with writing as the dependent variable and LD measures as the predictor variables, were performed to explore LD measures' ability to predict the IELTS 7 writing sublevels. Based on the simple count, *TTR* (F (1,41) = .103, p = .749, b = -.126, R^2 = .003) and *D* (F (1,41) = .598, p = .444, b = -.001, R^2 = .014) could not estimate the writing variances, and neither was the combination of these two measures significant. Similarly, for the flemma count, *TTR*

(*F* (1,41) = .039, *p* = .844, b = -.085, R^2 = .001) and *D* (*F* (1,41) = .331, *p* = .568, b = .001, R^2 = .008) could not significantly predict writing whether entered into the model separately or in combination. For the lemma count, *Guiraud's Index* (*F* (1,41) = .004, *p* = .950, b = .002, R^2 = .000) and *HD-D* (*F* (1,41) = .068, *p* = .796, b = -.139, R^2 = .002) were also not indicative of writing, either as separate or combined predictors. Additionally, the statistical power values were small for all these predictors: *TTR* (.06), and *D* (.09) for simple count, *TTR* (.05) and *D* (.07) for flemma count, and *Guiraud's Index* (.05) and *HD-D* (.06) for lemma count.

5.3.2.3 Exploring the extent to which LD measures predict IELTS-based writing proficiency 7.5 sub-levels (low and high) of L1 Chinese L2 English learners based on simple, flemma, and lemma counts.

Table 5.12 presents the median scores of the two (low and high) sublevels of IELTS writing 7.5 level and the observed G power values based on the simple, flemma, and lemma counts. The median scores were similar; however, the medians of the high-level group were higher than those of the low-level group for most LD measures. Cohen's *r* values indicated that the LD measure predictions of writing proficiency varied, depending on the analysis units. The simple count could best enhance the predictive powers of writing proficiency of all three basic measures (*Types, TTR, Guiraud's Index*). *D* and *HD-D* were more powerful writing indicators once the flemma count was used, whereas *MTLD* was better able to predict writing once the simple count was used. Interestingly, *TTR* was the most effective measure when both non-lemmatization and lemmatization methods were applied.

Friedman's Two-way ANOVA analyses were conducted to investigate whether the simple, flemma, and lemma counts were significantly different for all LD measures. All three analysis units differed significantly from each other for *TTR* and

D, but the flemma and lemma counts were not significantly different for *Types* or *Guiraud's Index*. The flemma count differed from the simple and lemma counts for *MTLD*. However, there were no significant differences between all three analysis units for *HD-D.*

Table 5.12

Medians and Cohen's r Values of Basic and Sophisticated LD Measures (IELTS 7.5 Level)

Basic LD Measures					Sophisticated LD Measures				
Measure	Low	High	*r*	Observed Power		Low	High	*r*	Observed Power
Types0	111.50	113.00	.006	.05	D0	63.20	61.45	.127	.36
Types1	104.00	106.50	.003	.05	D1	53.90	54.35	.143	.41
Types2	108.00	107.00	.013	.08	D2	57.15	56.30	.080	.30
TTR0	.56	.57	.017	.05	MTLD0	53.01	50.65	.053	.27
TTR1	.52	.54	.013	.12	MTLD1	50.99	50.28	.030	.22
TTR2	.54	.54	.017	.12	MTLD2	54.64	54.89	.040	.26
Guiraud0	7.89	7.99	.010	.05	HDD0	.77	.78	.187	.48
Guiraud1	7.35	7.54	.003	.05	HDD1	.77	.78	.191	.16
Guiraud2	7.64	7.57	.013	.07	HDD2	.78	.78	.050	.48

Table 5.13

Friedman's Two-way ANOVA Results of the Overall Differences Between the Simple, Flemma, and Lemma Counts (IELTS 7.5 Level)

Measures	$\chi 2$ (2)	*p*	0 vs 1	0 vs 2	1 vs 2
Types	45.187	<.001	*	*	NS
TTR	49.441	<.001	*	*	*
Guiraud	45.984	<.001	*	*	NS
D	49.750	<.001	*	*	*
MTLD	8.848	.012	*	NS	*
HDD	7.000	.030	NS	NS	NS

I further analyzed whether the LD measures were discriminative of IELTS writing 7 (low and high) sublevels using the Mann-Whitney U tests (see Table 5.14). The tests indicated similar findings to the IELTS 6.5 and 7 levels analyses in that none of the LD measures could discriminate between the low and high writing sub-groups of L1 Chinese writers at IELTS 7.5 level, based on all three analysis units. Among all LD measures, *HD-D* seemed more powerful in explaining the writing variances for the simple and flemma counts because of the smaller *p* values.

Table 5.14

Mann-Whitney U-test Results of the LD Measures' Discrimination Between IELTS 7.5 Writing Sub-Levels

Measure	Simple count			Flemma count			Lemma count		
	U	*p*	low - high	*U*	*p*	low - high	*U*	*p*	low - high
Types	127.00	.970	NS	127.50	.985	NS	126.00	.940	NS
TTR	125.50	.925	NS	126.00	.940	NS	125.50	.925	NS
Guiraud	126.50	.955	NS	127.50	.985	NS	126.00	.940	NS
D	109.00	.474	NS	106.50	.418	NS	116.00	.651	NS
MTLD	120.00	.763	NS	123.50	.865	NS	122.00	.821	NS
HD-D	100.00	.289	NS	99.50	.280	NS	120.50	.776	NS

The relationships between LD measures and writing were further examined using Spearman's correlation tests (see Table 5.15). The analyses show that the positive correlations between the LD measures ranged from low to high. As shown in Table 5.15, all LD measures had low positive correlations with writing. Using the data to identify the most powerful basic and sophisticated LD measure for each analysis unit, writing had the strongest correlations with *Guiraud's Index* and *HD-D* for the simple count, with *TTR* and *HD-D* for the flemma count, and with *TTR* and *D* for the lemma count.

Table 5.15

Correlations Between LD Measures and Writing (IELTS 7.5 Level)

		Types	TTR	Guiraud	D	MTLD	HDD
Simple count	Writing	.048	.043	.052	.232	.079	.287
Flemma count	Writing	.023	.035	.023	.241	.060	.275
Lemma count	Writing	.041	.044	.041	.147	.072	.134

To explore LD measures' discrimination between IELTS writing 7.5 sublevels, log-transformed regression analyses were conducted with the LD measures which were the most strongly related to writing for each analysis unit. For the simple count, though, neither *Guiraud's Index* (F (1,30) = .000, p = .983, b = .001, R^2 = .000) nor *HD-D* (F (1,30) = 2.454, p = .128, b = .756, R^2 = .076) were significant when separately analyzed. *Guiraud's Index* could not discern any writing variances, whereas *HD-D* could estimate 7.6% of the writing variances. However, the combination of these two measures turned significant (F (2,29) = 6.321, p = .005), and could explain 30.4% of the writing variances.

For the flemma count, *TTR* (F (1,30) = .006, p = .941, b = -.028) and *HD-D* (F = 2.500, p = .124, b = .801) could not support the regression model. *TTR* (R^2 = .000) could not significantly distinguish between low and high writing sublevels, whereas *HD-D* was more discriminative of writing (R^2 = .077). Furthermore, the combination of these two measures was significant (F = 5.212, p = .012) and could explain 26.4% of the writing variances. For the lemma count, neither *TTR* (F = .137, p = .714, b = .131, R^2 = .005) nor *D* (F = .727, p = .401, b = .001, R^2 = .024) were significant, either as separate or combined writing predictors (F = .690, p = .509, R^2 = .045). G power analysis also indicated that all these predictors received the small statistical power values for *Guiraud's Index* (.05) and *HD-D* (.31) based on simple

count, *TTR* (.05) and *HD-D* (.31) based on flemma count, and *TTR* (.07) and *D* (.13) based on lemma count.

5.4 Discussion

The current study investigated the extent to which LD measures predicted writing proficiency variation depending on L1 Chinese L2 English learners' writing proficiency. The study first descriptively summarized the writing variability within the three writing proficiency levels (6.5, 7, 7.5), considering the average scores and several dispersion scores, and I then compared the variabilities using Fisher's F test. Second, the study examined the extent to which LD measures predict writing proficiency variation depending on the L1 Chinese participants' writing proficiency levels.

First, the writing variability analyses showed that the data distribution patterns of the three writing proficiency (6.5, 7, 7.5) levels of L1 Chinese writers were similar. The mean scores of all three writing levels were consistent since the means became greater commensurate to the increasing proficiency levels. Even though the data sets had a similar distribution shape, they still might have had different variations. I, therefore, checked the data dispersions of each writing level. Several variability measures showed that the data dispersions of the three writing levels were small, implying that the writers at each writing level were less dispersed. Additional analyses of Fisher's two samples for variances revealed that the writing variations of 6.5, 7, and 7.5 levels were not significantly different from each other. These were the expected findings because of the small data sample sizes of the three writing levels (6.5 (N = 28), 7 (N = 43), and 7.5 (N = 32)).

Second, the effect sizes obtained from the Mann-Whitney U tests yielded clear findings of the analysis units' influence on LD measures' ability to predict writing.

The simple and flemma counts had a greater effect on LD measures' discriminative power of the IELTS 6.5 level's low and high sub-groups. *Types* and *Guiraud's Index* were more effective on flemmatized data. *TTR* and *MTLD* were more predictive of writing on non-lemmatized data, and *D* and *HD-D* were better writing discriminators on lemmatized data. They demonstrate that even within a single writing (6.5) level, the extent to which LD measures predict writing proficiency depends on the particular analysis unit used.

For IELTS writing 7, the simple and lemma counts were the more discriminating analysis units than the flemma count. The analyses showed that all LD measures except for *HD-D* were more reliable writing sub-level indicators when based on the simple count. This finding provides clear evidence of the greater appropriateness of the simple count over flemma, or lemma counts in discriminating between the IELTS writing 7 sub-levels (low and high).

For the 7.5 level, the simple and lemma counts could best enable and enhance the extent to which all three basic measures and *MTLD* predict writing proficiency. In contrast, the simple and flemma counts could make *D* and *HD-D* the more powerful measures. All three basic measures were more indicative of intra-group writing proficiency when a lemma count was used, and *TTR* and *MTLD* were stronger indicators when a simple count was used. *D* and *HD-D* were more effective in predicting writing when based on the flemma count.

Regarding the LD measures' discrimination between the low and high writing sub-levels, the Mann-Whitney U test results showed that none of the LD measures could significantly predict the low and high sub-levels of all three writing proficiency (6.5, 7, 7.5) levels, and that no analysis unit could significantly enhance their

predictive power of writing. Comparing the *p* values, though, with the smaller *p* values, *D* and *HD-D* seemed best able to differentiate between writing sub-groups.

Regression analyses indicated similar findings to those obtained with the Mann-Whitney U tests but also provided some additional findings. LD measures could not significantly contribute to the regression models. The sophisticated measures seemed more powerful writing (low and high) sublevel predictors than the basic measures for all three analysis units. More interestingly, the LD measures analyzed separately and in combination were not significant predictors of sublevels of IELTS writing 6.5 and 7 levels for all three analysis units or of the IELTS 7.5 sublevels for the lemma count. However, when combined, *Guiraud's Index* and *HD-D* for the simple count and *TTR* and *HD-D* for the flemma count could significantly discriminate between low and high writing sub-levels of IELTS 7.5.

In sum, for IELTS writing 6.5 sub-level differentiation, the three analysis units had different influences on the LD measures: for writing 7 level, I found the simple count to be the most impactful analysis; and for 7.5 level, the lemma count seemed to have the greatest influence on the LD measures. Thus, these reported findings of the analysis units' different influences on the LD measures, depending on the writing proficiency level, suggest that different analysis units should be adopted for different writing proficiency levels in studies examining LD measure ability to predict intra-group writing variations.

5.5 Limitations

Despite these useful new findings, the study includes at least one sizeable limitation that might influence the validity of the study's findings. This limitation relates to the small number of participants in each writing sub-groups (low and high). Since the study restricted the L1 background to only Chinese (the majority L1 group),

the data sample sizes were small, especially when further subdivided into two sub-groups within each of the three IELTS bands. The current study findings might thus be of limited generalizability and reliability. With this study on the language (writing) proficiency influence on the extent to which LD measures predict writing proficiency as a starting point, future research should conduct a wider investigation by using sufficiently large participant sample sizes and different L1 background groups.

5.6 Conclusion

This chapter has examined the extent to which LD measures predict writing proficiency was influenced by language proficiency (writing) and investigated the most discriminative word-counting criteria for each specific writing level (IELTS 6.5, 7, or 7.5). The current study's findings are restricted to only L1 Chinese L2 English learners because it controlled the L1 background.

The initial analysis of the writing variability showed that the writing scores of the participants within all three writing levels (6.5, 7, 7.5) were not widely dispersed. Regarding the LD measures' discrimination between writing sub-levels based on the three different analysis units, the findings showed that LD measures could not significantly predict IELTS 6.5, 7 or 7.5 low and high writing sub-groups using any of the three analysis units.

However, the combinations of *Guiraud's Index* and *HD-D* on the non-lemmatized data and *TTR* and *HD-D* on the flemmatized data were significant predictors of the IELTS 7.5 sublevels. For the IELTS 6.5 level, most LD measures were better writing sublevel discriminators when based on the simple and flemma counts, and each of the three analysis units had different effects on the extent to which LD measures predict writing proficiency. In predicting low and high sub-levels of IELTS 7 and 7.5, the LD measures were stronger writing indicators when based on

the simple and lemma counts. Five out of the six LD measures were best able to predict IELTS 7 writing sublevels once the simple count was used, whereas the lemma count could best increase the discriminative power of most LD measures in predicting the IELTS 7.5 sublevels.

The findings imply it seems inappropriate to assume that the smaller analysis unit (e.g., simple or lemma count) is always a more suitable unit for assessing the LD of low-proficiency L2 writers. The slightly larger analysis unit (e.g., flemma) should only sometimes be considered better for higher-proficiency writers. However, future research should conduct wider studies to examine the variation in the extent to which LD measures predict writing proficiency under the controlled text length based on different analysis units.

Chapter 6

Investigating the Minimum Constant Text Length Required for LD Measures to Predict Speaking Proficiency from Three Analysis Units

6.1 Introduction

Chapter 3 (writing data analysis) addressed recent controversy regarding the most appropriate analysis units to use in L2 vocabulary assessment by comparing the influences of different analysis units on the extent to which LD measures predict L2 writing proficiency. The study confirmed Treffers-Daller et al.'s (2018) findings that using the appropriate word-counting criteria is essential in LD assessment, and that the extent to which LD measures predict writing proficiency varies, depending on the analysis unit used. These findings have highlighted the necessity of exploring whether the word-counting criteria might also be a potential factor affecting the extent to which LD measures predict speaking proficiency, a different language mode being not as formal or complex as writing.

Yu (2010) compared the extent to which LD measures predict different language modes (writing and speaking). His study yielded the useful insights that written and spoken lexical diversity of the participants were at similar levels. However, he found that an LD measure (*D*) was a better speaking predictor (23.4% variance) than writing (11% variance). The findings implied that a single LD measure (*D*) indicated different predictive capacity for the two separate language modes (writing and speaking). Despite the examination of LD measures both as writing and speaking predictors, the findings were based on only a simple count because of the Michigan English Language Assessment Battery (MELAB) test's nature, that mainly demands learners' knowledge of word forms.

Most speaking proficiency tests (e.g., IELTS, TOEFL) require both L2 learners' inflectional and derivational knowledge. Therefore, the choice of the analysis unit that can accurately and actually capture learners' existing lexical knowledge is necessary in validating the extent to which LD measures predict speaking proficiency. As the literature review section highlighted, validation studies on the extent to which LD measures predict speaking proficiency (Read & Nation, 2006; Zhang & Daller, 2019) failed to investigate the influences of different analysis units on LD measures' estimation of the speaking variances. Specifically, no study to date has compared the simple, flemma, and lemma counts which variously calculate and represent different levels of learners' inflectional knowledge.

To address this deficiency, similar to the writing study (chapter 3), the current study (chapter 6) also partially replicates Treffers-Daller et al. (2018), which examined the extent to which LD measure predictions of general language proficiency based on simple, lemma, and word family counts. The current study investigates the influence of word-counting criteria on the extent to which LD measures predict speaking proficiency. For the comparability of the analysis unit influence on LD measures for different language proficiencies (general, writing, and speaking), this partial replication study adopted similar procedures to Treffers-Daller et al. (2018) and the writing data analysis in chapter 3.

Unlike the writing study, though, the current study could not address the influences of L1 background and L2 speaking proficiency on the extent to which LD measures predict speaking proficiency since the data sample was too small to identify different L1 backgrounds and different sub-level groups. However, the current study examined not only the influence of the analysis units but also an equally important factor: the influence of text length. The initial data analysis of the same cut-off point

(200 words) indicated a lack of predictive power for LD measures predictions of speaking proficiency, and so raised an important question to be further explored. By addressing this question, "What is the minimal constant spoken text length at which LD measures are predictive of speaking?", the current study examined different constant text lengths in increasing of 50 tokens (from 200 to 450 tokens) and also analyzed the full text lengths of varying token numbers to gain additional insights.

To investigate the extent to which LD measures predict L2 speaking proficiency and the minimum constant text length for the LD measures' greater predictions using simple, flemma, and lemma counts, the current chapter attempted to answer the following two research questions.

RQ 1: How do flemmatization and lemmatization influence LD scores and measures' discrimination between speaking proficiency levels (IELTS 6.5, 7, and 7.5) for different text lengths?

RQ 2: Based on different analysis units and spoken text lengths, to what extent do LD measures predict L2 speaking proficiency?

6.2 Method

6.2.1 Participants

The participants were L2 English learners enrolled in the 2021-2022 pre-sessional English program at a UK university. Out of the 171 students who gave access to their audio recordings via written consent form (see Appendix 1), I selected 68 spoken transcripts to match the written data analysis. The Research Ethics Committee of Queen Mary University (UK) has approved this research (approval number: QMREC2414a).

However, 13 transcripts were excluded since the transcripts could not be lemmatized. The participants in the current study (N = 55) were specializing in Humanities and Social Sciences, Law, and Science and Engineering. They were from 11 different L1 backgrounds: Chinese, Russian, Japanese, Indonesian, Colombian, and Germany made up most of the participants. Their IELTS-based speaking proficiencies ranged from 6.5 to 7.5, as shown in Table 6.1.

Table 6.1

Participants' IELTS-Based Speaking Proficiency

Speaking Levels	6.5	7	7.5	Total
N	17	19	19	55

6.2.2 Data and scoring

The data were recordings of the participants' seminar presentations, and each recording comprised two sections. In the first section, the participants presented on the same theme, "Globalization's impact on society", that had been used in the essays from which the written data was deprived (chapters 3, 4, and 5). In the second section, the participants answered their classmates' questions and discussed the ideas relating to their presentation. The class teacher assessed their speaking performance by using an IELTS-based speaking rubric (see Appendix 6) and classified the recordings into the different speaking (6.5, 7, 7.5) bands.

Initially, the audio files were transcribed using the Otter AI transcription tool (https://otter.ai). A cursory reading of the spoken transcripts of IELTS-based speaking levels 6.5, 7, and 7.5 (see Appendix 7) found that the discussion parts of the transcripts included more utterances by the classmates than the presenters. I therefore

analyzed only the presentation part of the transcripts; these only reflected the participant's vocabulary knowledge. The presentation text lengths ranged from 431 words to 1,437 words.

6.2.3 Data processing

Prior to analysis, I performed the data cleaning, flemmatizing, lemmatizing, and setting the constant text lengths. First, the questioners' or moderators' utterances and the presentation time markers were excluded from the spoken transcripts. Unlike for the written data, the spoken data cleaning required some additional steps (the removal of hesitation markers, repeats, backchannelling utterances, and false starts) since the data transcription might not have been fully accurate. Like other speech-to-text converters, the Otter software sometimes produced some unintelligible or inaccurate words because of misheard transcription due to some presenters' imperfect pronunciation. Therefore, the wrong words (e.g., "*there*" instead of "*they're*") were replaced with the right words when necessary. Following Treffers-Daller et al.'s (2018) study's procedure, proper names, cardinal numbers, and abbreviations were deleted to control LD score inflation. The contraction words (e.g., "*won't*", "*I'm*") were expanded into their equivalent full forms.

Second, Treffers-Daller et al. (2018) used the CLAN software to create different lemmatized text versions. However, I instead used the Python program in order to flemmatize and lemmatize both the written and spoken data. In manual editing, the flemmatized and lemmatized texts were checked for errors (e.g., failure to change the irregular comparative and superlative adjectives to the positive degree adjective).

Third, as the LD measures are sensitive to text sample size to differing degrees, comparing the LD scores of the same length texts has been recommended in

LD assessment (Treffers-Daller, 2013; Treffers-Daller et al., 2018). Therefore, the same token number (200 words) was taken from the middle of each presentation transcript using the Gramulator software.

6.2.4 Procedure

Since most prior LD studies have adopted different measures and different text lengths, as well as different statistical procedures, it sometimes appears hard to compare, validate, and generalize the findings. Therefore, for more reliable validation, similar to the written data analyses in the previous chapters, the current study also partially replicates the Treffers-Daller et al. (2018) study. I used the same procedures as for the written data, grounding it on Treffers-Daller et al.'s (2018) study but for a different language mode (speaking). For the analysis, three text (non-lemmatized, flemmatized, and lemmatized) versions of each presentation transcript were created, and 200 words (the same cut-off point as in Treffers-Daller et al.'s (2018) study and as in our written data analyses in chapters 3, 4, and 5) were taken from the middle of the spoken texts. Then the LD scores were computed, using the same six LD measures (*Types, TTR, Guiraud's Index, D, MTLD,* and *HD-D*).

However, the 200-word text analysis showed that the extent to which LD measures predict speaking proficiency was not significant. The finding clearly implied that the short constant text (200 words) length appeared insufficient for LD measures to predict speaking levels. I therefore expanded the study by exploring the constant text length(s) at which LD measures could be predictive of speaking proficiency. For this reason, I further analyzed different constant text lengths (250, 300, 350, 400, 450 words) as well as the full lengths (431 to 1,437 words) to explore the specific text length(s) at which LD measures indicated the ability to predict speaking proficiency. 55 spoken transcripts were analyzed for the 200- to 400-word

text analyses, but only 54 transcripts were analyzed for the 450-word text analysis since one spoken sample fell under the cut-off point (450 words). The following table shows the participant numbers at each of the IELTS-based speaking levels used in the different text length analyses.

Table 6.2

Participant Numbers at Each IELTS Speaking Level for Different Text Length Analyses

Speaking levels	Text Length						
	200	250	300	350	400	450	Full length
6.5	17	17	17	17	17	16	17
7	19	19	19	19	19	19	19
7.5	19	19	19	19	19	19	19
Total	55	55	55	55	55	54	55

6.2.5 Statistical analyses

First, a normality test, the Box Plot method, was used to examine the data distributions for the different text lengths. The findings indicated some data did not meet the normality assumption, including the mild outliers, and the full text length data included one extreme outlier. The data were not normally distributed and also each participant sample size (17, 19, 19) was too small to use the parametric tests. I therefore kept the outliers and decided to use non-parametric tests that are less sensitive to the skewed data from the outliers.

I examined the differences between the three analysis units by using the Friedman's Two-way ANOVA and LD measures' discrimination between the three speaking levels based on the three analysis units by using Kruskal-Wallis test. The

Bonferroni correction was used to adjust the alpha values for multiple comparisons. I also explored the correlations between the writing and LD score by using the Spearman's correlation test and performed the log transformed regression analyses to examine LD measures as the speaking predictors. I also performed the power analyses for Kruskal-Wallis and regression tests with G power software.

6.3 Results

The current partial replication study of Treffers-Daller et al. (2018) investigates the extent to which LD measures predict speaking proficiency depended on the analysis unit and the text sample size used. First, I reported instead the median scores of the mean scores because of the non-normally distributed data. The medians of the LD scores calculated on the non-lemmatized, flemmatized, and lemmatized spoken samples for varying text lengths (200, 250, 300, 350, 400, 450, and full) are presented in Table 6.3 (basic LD measures) and in Table 6.4 (sophisticated LD measures).

6.3.1 Flemmatization and lemmatization influences on LD scores and measures' discrimination between speaking proficiency levels (IELTS 6.5, 7, and 7.5) for different text lengths.

The median scores confirmed the expected findings of the different word-counting units' influence on LD scores. For all LD measures except *HD-D*, the LD scores were the highest on the raw data which had not been lemmatized, followed by the LD scores calculated on the lemmatized data, with the flemmatized data obtaining the lowest LD scores. However, the LD median scores on the three differently lemmatized text versions were similar for *HD-D*. These findings were the same across all different text lengths. Having looked at the LD median scores of each measure for

different speaking proficiency levels, the scores were not consistent across 6.5, 7, and 7.5 levels. For most LD measures, the highest (7.5) level speakers received the lowest LD scores. This finding was in line with the written data analyses in chapters 3, 4, and 5, whereas Treffers-Daller et al.'s (2018) study showed that higher proficiency writers used more diverse vocabulary.

Table 6.3

Median Values of Basic LD Measures for Different Text Lengths

Token	Level	Types0	Tyes1	Types2	TTR0	TTR1	TTR2	G0	G1	G2
200	6.5	112.00	104.00	105.00	.56	.52	.53	7.92	7.35	7.42
	7	118.00	109.00	112.00	.59	.55	.56	8.34	7.71	7.92
	7.5	105.00	101.00	103.00	.53	.51	.52	7.42	7.14	7.28
250	6.5	139.00	128.00	130.00	.56	.51	.52	8.79	8.10	8.22
	7	138.00	125.00	129.00	.55	.50	.52	8.73	7.91	8.16
	7.5	131.00	122.00	122.00	.52	.49	.49	8.29	7.72	7.72
300	6.5	157.00	143.00	144.00	.52	.48	.48	9.06	8.26	8.31
	7	157.00	143.00	146.00	.52	.48	.49	9.06	8.26	8.43
	7.5	150.00	140.00	142.00	.50	.47	.47	8.72	8.08	8.20
350	6.5	173.00	158.00	159.00	.49	.45	.45	9.25	8.45	8.50
	7	176.00	160.00	163.00	.50	.46	.47	9.41	8.55	8.71
	7.5	169.00	156.00	157.00	.48	.45	.45	9.03	8.34	8.39
400	6.5	189.00	174.00	174.00	.48	.44	.44	9.45	8.70	8.70
	7	194.00	177.00	179.00	.49	.44	.45	9.70	8.85	8.95
	7.5	181.00	171.00	172.00	.45	.43	.43	9.05	8.55	8.60
450	6.5	210.00	192.50	197.00	.47	.43	.44	9.90	9.08	9.29
	7	213.00	192.00	192.00	.47	.43	.43	10.04	9.05	9.05
	7.5	196.50	180.50	183.50	.44	.41	.41	9.36	8.58	8.72
Full	6.5	262.00	234.00	242.00	.42	.38	.39	10.41	9.72	9.45
	7	303.00	268.00	280.00	.40	.37	.37	10.77	9.42	9.68
	7.5	255.00	233.00	236.00	.38	.33	.34	9.81	8.80	8.95

Table 6.4

Median Values of Sophisticated LD Measures for Different Text Lengths

Token	Level	D0	D1	D2	MTLD0	MTLD1	MTLD2	HDD0	HDD1	HDD2
200	6.5	65.90	53.30	55.50	51.45	51.93	52.01	.78	.77	.77
	7	72.00	57.50	60.40	55.84	56.12	58.33	.79	.78	.79
	7.5	59.90	53.40	54.80	56.30	58.04	55.13	.78	.77	.76
250	6.5	67.20	52.30	53.40	53.46	52.59	51.14	.77	.77	.77
	7	69.00	56.20	58.20	56.84	54.76	56.89	.78	.79	.78
	7.5	63.90	51.20	55.10	57.07	53.44	54.28	.77	.77	.78
300	6.5	68.30	54.10	55.90	50.73	48.21	54.08	.78	.78	.78
	7	66.10	53.00	57.70	55.97	55.02	54.90	.78	.78	.78
	7.5	64.60	57.50	57.80	52.27	51.30	54.60	.78	.78	.78
350	6.5	64.70	53.90	56.60	51.50	51.01	52.07	.78	.77	.78
	7	65.30	57.40	58.40	54.77	57.89	54.38	.78	.78	.78
	7.5	69.20	58.50	58.80	55.83	52.60	54.53	.79	.78	.79
400	6.5	69.50	54.40	55.40	52.18	49.06	51.86	.78	.78	.78
	7	70.80	57.60	60.40	56.35	53.57	56.46	.79	.78	.79
	7.5	65.80	56.90	56.70	51.49	52.77	51.19	.78	.78	.78
450	6.5	69.45	56.60	58.20	52.21	50.40	51.33	.78	.78	.78
	7	72.30	58.70	63.40	57.53	56.37	57.06	.79	.79	.79
	7.5	66.30	58.55	57.75	54.37	51.52	52.32	.78	.78	.78
Full	6.5	72.00	57.70	57.60	52.82	48.99	52.40	.79	.78	.79
	7	74.90	61.50	63.20	59.15	56.01	56.41	.80	.79	.79
	7.5	70.10	55.80	57.60	52.21	47.38	49.79	.79	.78	.79

I statistically examined the differences between the three analysis units for all different text lengths by using the Friedman's Two-way ANOVA analyses. Tables 6.5 and 6.6 indicate that simple, flemma, and lemma counts differed significantly from each other for all three basic measures (*Types, TTR, Guiraud's Index*) across all different constant text lengths. When the whole spoken transcripts of varying lengths were examined, though, the flemma and lemma counts were not significantly different for *TTR*.

As for the sophisticated LD measures, there were significant differences between all three analysis units for *D* on 200, 250, 300, 450, and full lengths, whereas the flemma and lemma counts were not significantly different for the 350 and 400 tokens. For *MTLD*, the simple and lemma counts were significantly different on 200-word texts, the simple and flemma counts were different on 350, 400, and 450 tokens, and all three units were different on the full length texts. For *HD-D*, the simple count differed significantly from the flemma and lemma counts on the full length texts, the simple count and lemma counts were significantly different on 200-word texts, and the simple and flemma counts were significantly different on 300- and 400-word texts.

Table 6.5

Overall Differences Between the Simple, Flemma, and Lemma Counts (200, 250, 300, 350 Text Lengths)

Token		Friedman's Two-way ANOVA					
		Types	TTR	Guiraud	D	MTLD	HD-D
200	$\chi 2$	103.72	103.19	103.42	90.97	9.37	22.67
	p	.000	.000	.000	.000	.009	<.001
	0 vs 1	*	*	*	*	NS	NS
	0 vs 2	*	*	*	*	*	*
	1 vs 2	*	*	*	*	NS	NS
250	$\chi 2$	99.47	100.04	99.47	89.05	3.964	5.826
	p	.000	.000	.000	.000	.138	.054
	0 vs 1	*	*	*	*	NS	NS
	0 vs 2	*	*	*	*	NS	NS
	1 vs 2	*	*	*	*	NS	NS
300	$\chi 2$	93.30	100.01	97.170	92.40	3.38	18.85
	p	.000	.000	.000	.000	.184	<.001
	0 vs 1	*	*	*	*	NS	*
	0 vs 2	*	*	*	*	NS	NS
	1 vs 2	*	*	*	*	NS	NS
350	$\chi 2$	82.30	81.09	82.30	67.95	10.35	10.62
	p	.000	.000	.000	<.001	.006	.005
	0 vs 1	*	*	*	*	*	NS
	0 vs 2	*	*	*	*	NS	NS
	1 vs 2	*	*	*	NS	NS	NS

Table 6.6

Overall Differences Between the Simple, Flemma, and Lemma Counts (400, 450, Full Lengths)

Tokens	Friedman's Two-way ANOVA						
		Types	TTR	Guiraud	D	MTLD	HD-D
400	$\chi 2$	99.90	99.25	99.90	88.11	9.86	23.55
	p	.000	.000	.000	.000	.007	<.001
	0 vs 1	*	*	*	*	*	*
	0 vs 2	*	*	*	*	NS	NS
	1 vs 2	*	*	*	NS	*	NS
450	$\chi 2$	103.09	102.18	103.09	88.26	8.926	18.97
	p	.000	.000	.000	.000	.012	<.001
	0 vs 1	*	*	*	*	*	NS
	0 vs 2	*	*	*	*	NS	NS
	1 vs 2	*	*	*	*	NS	NS
Full	$\chi 2$	109.03	100.09	98.76	90.15	38.06	47.09
	p	.000	.000	.000	.000	<.001	<.001
	0 vs 1	*	*	*	*	*	*
	0 vs 2	*	*	*	*	*	*
	1 vs 2	*	NS	*	*	*	NS

Epsilon squared values showed the different influences of the three analysis units on LD measures' discrimination between different speaking proficiency levels based on different text lengths (see Table 6.7). Among the three analysis units, the simple count had the greatest impact on all three basic measures (*Types, TTR, Guiraud's Index*), but it was the least impactful unit on *MTLD* and *HD-D*. The

flemma and lemma counts had greater influences on *MTLD,* with the flemma count also being the most effective unit for *HD-D*. With the higher Epsilon values, the basic LD measures seemed more discriminative of speaking levels than the sophisticated LD measures.

Based on the simple count, *Types* could best discriminate between speaking levels on all different text lengths except for 350- and 450-word texts, on which the lemma count was the most discriminating unit. *TTR* could discern more speaking variances on 200, 250, 400, and 450 tokens when a simple count was used, and it was also a better discriminator of speaking levels on 300- and 350-word lemmatized texts. The flemma count could best enable and enhance *TTR*'s discriminative power of speaking levels on the full length texts. Based on the simple count, *Guiraud's Index* could discern the largest proportion of the speaking variances on all different text lengths except for 350 tokens, on which the lemma count was the most discriminating unit.

As for the sophisticated measures, *D* was the best speaking discriminator on the non-lemmatized data for the 200-, 250-, 300-, and 450-word, and full length texts and on the lemmatized data for 300, 350, and 400 tokens. The flemma count was the least impactful unit and could not even discern any speaking variances on 300-word texts. The flemma and lemma counts had greater effects on *MTLD's* discrimination of speaking levels than the simple count. *MTLD* could best distinguish between speaking levels on non-lemmatized full length texts, on 250-, 300-, and 450-word flemmatized texts and on 200-, 350-, and 400-word lemmatized texts.

Table 6.7

Three Analysis Units' Influences on LD Measures' Discrimination Between Different Speaking Levels for Different Text Lengths

Measure	Kruskal-Wallis Tests Epsilon Squared							Observed power values						
	200	250	300	350	400	450	Full length	200	250	300	350	400	450	Full length
Types0	.049	.054	.083	.030	.104	.115	.090	.31	.35	.42	.20	.61	.67	.56
Types1	.019	.033	.045	.065	.083	.107	.082	.16	.23	.38	.35	.49	.61	.56
Types2	.025	.035	.066	.080	.087	.116	.089	.15	.24	.34	.40	.55	.66	.56
TTR0	.040	.062	.067	.023	.114	.130	.124	.21	.44	.21	.32	.63	.63	.75
TTR1	.032	.037	.043	.070	.082	.104	.160	.21	.21	.21	.44	.50	.63	.75
TTR2	.024	.033	.068	.089	.094	.126	.132	.22	.34	.21	.44	.63	.63	.75
Guiraud0	.045	.053	.065	.030	.104	.114	.105	.29	.34	.42	.21	.62	.67	.45
Guiraud1	.021	.033	.045	.065	.083	.107	.104	.17	.23	.37	.35	.50	.69	.56
Guiraud2	.025	.036	.059	.080	.087	.114	.100	.14	.24	.38	.41	.54	.65	.46
D0	.013	.013	.003	.004	.023	.034	.030	.06	.06	.11	.07	.16	.20	.22
D1	.006	.006	.000	.005	.020	.023	.029	.54	.06	.11	.08	.14	.15	.21
D2	.010	.005	.003	.011	.025	.028	.028	.07	.06	.12	.10	.16	.17	.19
MTLD0	.001	.004	.008	.011	.032	.054	.073	.06	.05	.10	.07	.10	.27	.31
MTLD1	.000	.008	.015	.036	.027	.067	.066	.05	.05	.12	.12	.14	.31	.34
MTLD2	.005	.005	.012	.016	.047	.050	.057	.06	.05	.10	.10	.17	.25	.26
HD-D0	.006	.006	.003	.007	.036	.034	.030	.16	.16	.99	.05	.32	.32	.30
HD-D1	.007	.010	.003	.013	.021	.035	.024	.05	.16	.97	.16	.32	.32	.30
HD-D2	.016	.008	.004	.006	.028	.028	.022	.11	.16	.99	.16	.77	.05	.30

Kruskal-Wallis tests were performed to explore the LD measures' discriminative power of speaking levels, based on different analysis units, across all different text lengths. The basic and sophisticated LD measures' discrimination of speaking levels for the simple count is presented in Tables 6.8 and 6.9, for the flemma count in Tables 6.10 and 6.11, and for the lemma count in Tables 6.12 and 6.13.

The findings show that the LD measures seemed more discriminative of speaking levels as the texts became longer. Among all six LD measures, only *TTR* could significantly discriminate between the lowest (6.5) and the highest (7.5) speaking levels on full length texts with varying tokens (431 - 1437), based on all three analysis units. The other LD measures were not significant speaking discriminators, regardless of the analysis units and text lengths; however, the measures were found to become more powerful in discriminating speaking levels as the text lengths increased since the *p* values were getting smaller with the longer text lengths. In particular, basic LD measures (*Types, TTR, Guiraud's Index*) tended to be significant from 400 tokens onwards when simple and lemma counts were used. For the flemma count, the basic LD measures indicated the tendency to be significant on 450-word and full length texts. The *H* values showed it was clear that the basic measures seemed more powerful in discriminating between speaking proficiencies than the sophisticated measures for all three analysis units.

Table 6.8

Kruskal-Wallis Tests and Basic LD Measures' Discrimination Between Speaking Levels for Different Text Lengths (Simple Count)

Tokens	Measure	*H (2)*	*p*	6.5 vs 7	6.5 vs 7.5	7 vs 7.5
200	Types	2.632	.268	NS	NS	NS
	TTR	1.701	.427	NS	NS	NS
	Guiraud	2.407	.300	NS	NS	NS
250	Types	2.900	.235	NS	NS	NS
	TTR	3.372	.185	NS	NS	NS
	Guiraud	2.876	.237	NS	NS	NS
300	Types	4.467	.107	NS	NS	NS
	TTR	3.621	.164	NS	NS	NS
	Guiraud	3.215	.172	NS	NS	NS
350	Types	1.610	.447	NS	NS	NS
	TTR	1.240	.538	NS	NS	NS
	Guiraud	1.610	.447	NS	NS	NS
400	Types	5.620	.060	NS	NS	NS
	TTR	6.177	.046	NS	NS	NS
	Guiraud	5.620	.060	NS	NS	NS
450	Types	6.083	.048	NS	NS	NS
	TTR	6.914	.032	NS	NS	**NS**
	Guiraud	6.038	.049	NS	**NS**	NS
Full length	Types	4.863	.088	NS	NS	NS
	TTR	6.714	.035	NS	*	NS
	Guiraud	5.648	.059	NS	NS	NS

Table 6.9

Kruskal-Wallis Tests and Sophisticated LD Measures' Discrimination Between Speaking Levels for Different Text Lengths (Simple Count)

Tokens	Measure	*H (2)*	*p*	6.5 vs 7	6.5 vs 7.5	7 vs 7.5
200	D	.723	.697	NS	NS	NS
	MTLD	.040	.980	NS	NS	NS
	HD-D	.337	.845	NS	NS	NS
250	D	.716	.699	NS	NS	NS
	MTLD	.232	.891	NS	NS	NS
	HD-D	.314	.855	NS	NS	NS
300	D	.141	.932	NS	NS	NS
	MTLD	.441	.802	NS	NS	NS
	HD-D	.159	.924	NS	NS	NS
350	D	.217	.897	NS	NS	NS
	MTLD	.617	.735	NS	NS	NS
	HD-D	.404	.817	NS	NS	NS
400	D	1.226	.542	NS	NS	NS
	MTLD	1.731	.421	NS	NS	NS
	HD-D	1.963	.375	NS	NS	NS
450	D	1.820	.403	NS	NS	NS
	MTLD	2.836	.242	NS	NS	NS
	HD-D	1.823	.402	NS	NS	NS
Full length	D	1.638	.441	NS	NS	NS
	MTLD	3.933	.140	NS	NS	NS
	HD-D	1.632	.442	NS	NS	NS

Table 6.10

Kruskal-Wallis Tests and Basic LD Measures' Discrimination Between Speaking Levels for Different Text Lengths (Flemma Count)

Tokens	Measure	*H (2)*	*p*	6.5 vs 7	6.5 vs 7.5	7 vs 7.5
200	Types	1.048	.592	NS	NS	NS
	TTR	1.701	.427	NS	NS	NS
	Guiraud	1.151	.562	NS	NS	NS
250	Types	1.779	.411	NS	NS	NS
	TTR	1.998	.368	NS	NS	NS
	Guiraud	1.779	.411	NS	NS	NS
300	Types	2.412	.299	NS	NS	NS
	TTR	2.338	.311	NS	NS	NS
	Guiraud	2.412	.299	NS	NS	NS
350	Types	3.525	.172	NS	NS	NS
	TTR	3.766	.152	NS	NS	NS
	Guiraud	3.525	.172	NS	NS	NS
400	Types	4.494	.106	NS	NS	NS
	TTR	4.446	.108	NS	NS	NS
	Guiraud	4.494	.106	NS	NS	NS
450	Types	5.691	.058	NS	NS	NS
	TTR	5.512	.064	NS	NS	NS
	Guiraud	5.691	.058	NS	NS	NS
Full length	Types	4.439	.109	NS	NS	NS
	TTR	8.655	.013	NS	*	NS
	Guiraud	5.641	.060	NS	NS	NS

Table 6.11

Kruskal-Wallis Tests for Sophisticated LD Measures' Discrimination Between Speaking Levels for Different Text Lengths (Flemma Count)

Tokens	Measure	*H (2)*	*p*	6.5 vs 7	6.5 vs 7.5	7 vs 7.5
200	D	.311	.856	NS	NS	NS
	MTLD	.011	.994	NS	NS	NS
	HD-D	.366	.833	NS	NS	NS
250	D	.338	.844	NS	NS	NS
	MTLD	.407	.816	NS	NS	NS
	HD-D	.530	.767	NS	NS	NS
300	D	.151	.927	NS	NS	NS
	MTLD	.788	.674	NS	NS	NS
	HD-D	.140	.932	NS	NS	NS
350	D	.279	.870	NS	NS	NS
	MTLD	1.965	.374	NS	NS	NS
	HD-D	.711	.701	NS	NS	NS
400	D	1.079	.583	NS	NS	NS
	MTLD	1.476	.478	NS	NS	NS
	HD-D	1.113	.573	NS	NS	NS
450	D	1.216	.544	NS	NS	NS
	MTLD	3.555	.169	NS	NS	NS
	HD-D	1.879	.391	NS	NS	NS
Full length	D	1.545	.462	NS	NS	NS
	MTLD	3.564	.168	NS	NS	NS
	HD-D	1.278	.528	NS	NS	NS

Table 6.12

Kruskal-Wallis Tests for Basic LD Measures' Discrimination Between Speaking Levels for Different Text Lengths (Lemma Count)

Tokens	Measure	*H (2)*	*p*	6.5 vs 7	6.5 vs 7.5	7 vs 7.5
200	Types	1.373	.503	NS	NS	NS
	TTR	1.290	.525	NS	NS	NS
	Guiraud	1.373	.503	NS	NS	NS
250	Types	1.910	.385	NS	NS	NS
	TTR	1.781	.410	NS	NS	NS
	Guiraud	1.910	.385	NS	NS	NS
300	Types	3.542	.170	NS	NS	NS
	TTR	3.687	.158	NS	NS	NS
	Guiraud	3.181	.204	NS	NS	NS
350	Types	4.340	.114	NS	NS	NS
	TTR	4.787	.091	NS	NS	NS
	Guiraud	4.340	.114	NS	NS	NS
400	Types	4.678	.096	NS	NS	NS
	TTR	5.047	.080	NS	NS	NS
	Guiraud	4.678	.096	NS	NS	NS
450	Types	6.148	.046	NS	NS	NS
	TTR	6.697	.035	NS	NS	NS
	Guiraud	6.055	.048	NS	NS	NS
Full length	Types	4.798	.091	NS	NS	NS
	TTR	7.125	.028	NS	*	NS
	Guiraud	5.383	.068	NS	NS	NS

Table 6.13

Kruskal-Wallis Tests for Sophisticated LD Measures' Discrimination Between Speaking Levels for Different Text Lengths (Lemma Count)

Tokens	Measure	*H (2)*	*p*	6.5 vs 7	6.5 vs 7.5	7 vs 7.5
200	D	.553	.758	NS	NS	NS
	MTLD	.270	.874	NS	NS	NS
	HD-D	.856	.652	NS	NS	NS
250	D	.248	.883	NS	NS	NS
	MTLD	.253	.881	NS	NS	NS
	HD-D	.405	.817	NS	NS	NS
300	D	.151	.927	NS	NS	NS
	MTLD	.667	.717	NS	NS	NS
	HD-D	.201	.905	NS	NS	NS
350	D	.604	.739	NS	NS	NS
	MTLD	.839	.657	NS	NS	NS
	HD-D	.318	.853	NS	NS	NS
400	D	1.325	.516	NS	NS	NS
	MTLD	2.546	.280	NS	NS	NS
	HD-D	1.526	.466	NS	NS	NS
450	D	1.477	.478	NS	NS	NS
	MTLD	2.627	.269	NS	NS	NS
	HD-D	1.468	.480	NS	NS	NS
Full length	D	1.494	.474	NS	NS	NS
	MTLD	3.067	.216	NS	NS	NS
	HD-D	1.198	.549	NS	NS	NS

6.3.2 Exploring the extent to which LD measures predict L2 speaking proficiency for different analysis units and text lengths.

I conducted a series of correlation and regression analyses to investigate LD measures as speaking predictors based on the simple, flemma, and lemma counts across different text lengths, First, I investigated the relationships between the LD measures and speaking proficiency by using Spearman's correlation tests. As expected, the LD measures were strongly and positively correlated with each other for all different analysis units and text lengths, as all are indeed designed to capture the same construct (lexical diversity). However, speaking showed low negative correlations with the basic LD measures but low positive correlations with the sophisticated LD measures.

Table 6.14 shows that speaking was negatively correlated with all three basic measures on all different text lengths except for the full length, on which *Types* showed a positive correlation with speaking based on the simple count. All three sophisticated LD measures indicated the low positive correlations with speaking on all text lengths except for *MTLD*, which had no correlation with speaking on 450-word texts. For the basic measures, speaking was most strongly correlated with *Types* on 200-, 300-, and 350-word texts, with *TTR* strongest on 250, 400-, 450-word and full length texts, and with *Guiraud's Index* strongest on 350-word texts. For the sophisticated measures, speaking had the highest correlations with *D* on 350- and 450-word texts, with *MTLD* strongest on 200-word texts, and with *HD-D* on 200-, 250-, 300-, 400-word and full-length texts.

Table 6.14

Correlations Between LD measures and Speaking for Different Text Lengths (Simple Count)

Measure	Speaking						
	200	250	300	350	400	450	Full length
Types	-.069	-.097	-.164	-.094	-.192	-.215	.044
TTR	-.042	-.112	-.139	-.079	-.212	-.235	-.307*
Guiraud	-.065	-.096	-.138	-.094	-.192	-.213	-.110
D	.015	.038	.078	.139	.078	.048	.072
MTLD	.034	.025	.099	.087	.086	.000	.010
HD-D	.034	.068	.113	.117	.087	.047	.117

Table 6.15 indicates that there were low negative correlations between the basic LD measures and speaking based on the flemma count, but low positive correlations between the sophisticated measures and speaking on all different spoken sample sizes. However, *Types* was positively correlated with speaking on 200-word and full length texts, and *MTLD* showed negative correlation with speaking on 450-word and full length texts. For basic measures, the strongest correlations existed between *TTR* and speaking on 200-, 250-, 350-, 400-word and full-length texts while *Types* and *Guiraud's Index* were the most strongly correlated with speaking on 350- and 450-word texts. Among sophisticated measures, *D* had the strongest correlation with speaking on 200-, 300-, 350- and 400-word texts, whereas *HD-D* was most strongly related with speaking on 250-, 450-word and full-length texts.

Table 6.15

Correlations Between LD measures and Speaking for Different Text Lengths (Flemma Count)

Measure	Speaking						
	200	250	300	350	400	450	Full length
Types	.005	-.060	-.088	-.134	-.165	-.214	.038
TTR	-.026	-.066	-.083	-.136	-.177	-.208	-.358**
Guiraud	-.005	-.060	-.088	-.134	-.165	-.214	-.175
D	.052	.066	.119	.123	.101	.045	.061
MTLD	.029	.017	.092	.047	.041	-.051	-.029
HD-D	.050	.069	.097	.086	.090	.053	.090

Table 6.16 reports the findings of the relationships between LD measures and speaking, based on the lemma count. The basic LD measures, except for *Types* on full length texts, were negatively correlated with speaking, whereas the sophisticated measures, except *MTLD* on 450-word and full length texts, were positively correlated with speaking. Similar to the simple and flemma count analyses, the correlations between LD measures and speaking were low. For the basic measures, *TTR* had the strongest relationship with speaking on all different length texts except 250-word texts while *Types* and *Guiraud's Index* were most strongly related to speaking on 250-word texts. As for the sophisticated measures, speaking had the strongest correlations with *D* on 300-, 350-word and full-length texts, with *MTLD* on 200- and 450-word texts, and with *HD-D* on 250- and 400-word texts.

Table 6.16

Correlations Between LD Measures and Speaking for Different Text Lengths (Lemma Count)

Measure	Speaking						
	200	250	300	350	400	450	Full length
Types	-.020	-.084	-.133	-.164	-.184	-.232	.021
TTR	-.023	-.079	-.139	-.175	-.199	-.252	-.321*
Guiraud	-.020	-.084	-.115	-.164	-.184	-.228	-.155
D	.006	.066	.103	.076	.083	.039	.066
MTLD	.026	.035	.038	.005	.052	-.074	-.010
HD-D	.022	.073	.093	.095	.092	.052	.059

Second, I further explored the extent to which LD measures predict different speaking proficiency levels (6.5, 7, 7.5) varies, depending on whether simple (Table 6.17), flemma (Table 6.18), and lemma (Table 6.19) counts are used for differing text lengths. Log-transformed regression analyses were performed with the one basic and the one sophisticated LD measure which were most strongly correlated with speaking both as separate and combined indicators of speaking proficiency. When two basic or sophisticated measures indicated the same degrees of correlation with speaking, I only used the one measure which obtained the higher *H* values in Kruskal-Wallis tests. I also report statistical power values in Tables 6.17, 6.18, and 6.19.

Table 6.17 shows the findings of LD measures' ability to predict speaking proficiency levels, based on the simple count for different text lengths. *Types* and *HD-D* on 200-word texts, and *TTR* and *HD-D* on 250-word texts, were not effective either as separate or combined speaking predictors. However, *Types* and *HD-D* on 300-word

texts, and *Types* and *D* on 350-word texts, tended to be more powerful indicators once combined. For the longer text lengths, the combination of *TTR* and *HDD* on 400-word texts, and *TTR* and *D* on 450-word texts, could significantly discern the speaking variances. For full length texts of varying tokens, *TTR* ($F = 4.823$, $p = .032$, b = -1.681, $R^2 = .083$) was a significant speaking predictor, and the combined model of *TTR* and *HD-D* could predict 21.2% of speaking variances with both measures also being significant predictors when used separately. For most text lengths, with their higher *F* values, the basic LD measures appeared to be the stronger indicators of speaking proficiency than the sophisticated measures.

Table 6.18 shows whether LD measures were predictive of speaking levels, based on the flemma count. For the short lengths (200, 250, 300 tokens), *TTR* and *D* on 200-word texts, *TTR* and *HD-D* on 250-word texts, and *Types* and *D* on 300-word texts could not predict speaking levels when analyzed either separately or in combination. For 350- and 450-word texts, *TTR* and *D* indicated significant speaking predictions once the measures were entered into the model combined. Similarly, *Types* and *HDD* were significant speaking predictors once combined. For full length texts, *TTR* ($F = 6.023$, $p = .017$, b = -1.941) could discern 10.2% of the speaking variances as a separate measure, and *TTR* and *HD-D* were significant speaking predictors when combined.

Table 19 presents the findings regarding LD measures as speaking predictors, based on the lemma count. The short length text analyses yielded similar findings to the simple- and flemma-based analyses in that the measures under investigation were not effective in predicting speaking levels. *TTR* and *MTLD* on 200-word texts, *Types* and *HD-D* on 250-word texts, and *TTR* and *D* on 300-word texts were not significant separate speaking predictors, and neither were the combined models for these three

shorter text lengths. On 300- and 350-word texts, *TTR* and *HD-D* were reliable speaking predictors once combined. However, on 450-word texts, neither *TTR* nor *MTLD* could discern any speaking variances, either as separate or combined measures. On full length texts, *TTR* ($F = 5.495$, $p = .023$, b = -1.870, $R^2 = .094$) was significant both as a separate measure and when combined with *D*.

Table 6.17

Regression Results Reporting the Extent to which LD Measures Predict Speaking Proficiency (Simple Count)

Length	Entry	Speaking predictors	*F*	*p*	b	R^2	Observed power
200	Separate	Types	.433	.513	-.003	.008	.10
		HD-D	.017	.897	.166	.000	.05
	Combined	Types + HD-D	.712	.495		.027	
		Types		.241	-.008		
		HD-D		.324	1.981		
250	Separate	TTR	.873	.354	-.837	.016	.15
		HD-D	.085	.771	.420	.002	.06
	Combined	TTR + HD-D	1.686	.195		.061	
		TTR		.076	-2.469		
		HD-D		.122	3.419		
300	Separate	Types	1.251	.268	-.003	.023	.20
		HD-D	.503	.481	1.114	.009	.11
	Combined	Types + HD-D	3.146	.051		.108	
		Types		.020	-.010		
		HD-D		.030	4.758		
350	Separate	Types	.548	.462	-.002	.101	.68
		D	.961	.331	.003	.018	.17
	Combined	Types + D	2.648	.080		.092	
		Types		.044	-.008		
		D		.035	.010		
400	Separate	TTR	2.290	.136	-1.448	.041	.33
		HDD	.254	.616	.879	.005	.08
	Combined	TTR + HD-D	3.988	.024		.133	
		TTR		.008	-3.615		
		HD-D		.023	5.467		
450	Separate	TTR	2.812	.100	-1.590	.051	.39
		D	.262	.611	.002	.005	.08
	Combined	TTR + D	4.334	.018		.145	
		TTR		.006	-3.586		
		D		.022	.010		
Full length	Separate	TTR	4.823	.032	-1.681	.083	.59
		HD-D	.789	.378	1.782	.015	.15
	Combined	TTR + HD-D	6.995	.002		.212	
		TTR		<.001	-3.193		
		HD-D		.005	6.515		

Table 6.18

Regression Results Reporting the Extent to which LD Measures Predict Speaking Proficiency (Flemma Count)

Length	Entry	Speaking predictors	*F*	*p*	b	R^2	Observed power
200	Separate	TTR	.112	.739	-.292	.002	.06
		D	.133	.717	.001	.002	.06
	Combined	TTR + D	.618	.543		.023	
		TTR		.298	-1.530		
		D		.294	.006		
250	Separate	TTR	.385	.538	-.567	.007	.09
		HD-D	.042	.839	.304	.001	.06
	Combined	TTR + HD-D	.793	.458		.030	
		TTR		.220	-1.804		
		HD-D		.278	2.589		
300	Separate	Types	.453	.504	-.002	.008	.10
		D	.741	.393	.003	.014	.14
	Combined	Types + D	2.386	.102		.084	
		Types		.051	-.009		
		D		.043	.012		
350	Separate	TTR	1.033	.314	-.986	.019	.17
		D	.894	.349	.004	.017	.16
	Combined	TTR + D	3.762	.030		.126	
		TTR		.014	-3.393		
		D		.015	.015		
400	Separate	TTR	1.931	.170	-1.407	.035	.28
		D	.454	.504	.003	.008	.10
	Combined	TTR + D	4.081	.023		.136	
		TTR		.008	-3.779		
		D		.017	.014		
450	Separate	Types	2.488	.121	-.003	.046	.36
		HD-D	.063	.803	.433	.001	.06
	Combined	Types + HD-D	3.581	.035		.123	
		Types		.010	-.008		
		HD-D		.039	5.060		
Full length	Separate	TTR	6.023	.017	-1.941	.102	.69
		HD-D	.760	.387	1.686	.014	.14
	Combined	TTR + HD-D	8.247	<.001		.241	
		TTR		<.001	-3.548		
		HD-D		.003	6.479		

Table 6.19

Regression Results Reporting the Extent to which LD Measures Predict Speaking Proficiency (Lemma Count)

Length	Entry	Speaking predictors	*F*	***p***	b	R^2	Observed power
200	Separate	TTR	.072	.789	-.231	.001	.06
		MTLD	.001	.979	.000	.000	.05
	Combined	TTR + MTLD	.087	.917		.003	
		TTR		.679	-.529		
		MTLD		.749	.001		
250	Separate	Types	.610	.438	-.003	.011	.12
		HD-D	.079	.780	.406	.001	.06
	Combined	Types + HD-D	1.268	.290	-.009	.047	
		Types		.123	-.009		
		HD-D		.172	3.081		
300	Separate	TTR	1.091	.301	-.978	.020	.18
		D	.328	.569	.002	.006	.09
	Combined	TTR + D	2.796	.070		.097	
		TTR		.026	-3.122		
		D		.040	.012		
350	Separate	TTR	1.466	.231	-1.168	.027	.23
		HD-D	.356	.553	1.004	.007	.09
	Combined	TTR + HD-D	3.375	.042		.115	
		TTR		.015	-3.441		
		HD-D		.027	5.354		
400	Separate	TTR	2.270	.138	-1.468	.041	.33
		HD-D	.217	.643	.817	.004	.07
	Combined	TTR + HD-D	4.165	.021		.138	
		TTR		.006	-3.891		
		HD-D		.019	5.849		
450	Separate	TTR	3.143	.082	-1.688	.239	.98
		MTLD	.278	.600	-.002	.005	.08
	Combined	TTR + MTLD	1.993	.147		.072	
		TTR		.060	-2.515		
		MTLD		.360	.005		
Full length	Separate	TTR	5.495	.023	-1.870	.094	.65
		D	.536	.467	.003	.010	.11
	Combined	TTR + D	7.878	.001		.233	
		TTR		<.001	-3.680		
		D		.003	.015		

Overall, the LD measures were found to be not significant speaking predictors on the short length texts (200, 250, 300, 350 tokens), but the measures became more powerful from 400 words onwards. Only *TTR* was effective as a separate standalone speaking indicator when the texts of varying tokens were examined, regardless of the analysis unit used, and *TTR*'s predictive power was the strongest based on the simple count. Their higher *F* values implied a greater speaking predictive power of the basic LD measures than the sophisticated LD measures for most text lengths.

6.4 Discussion

The current study partially replicated Treffers-Daller et al. (2018), who examined the written lexical diversity role in predicting general language proficiency levels based on three analysis units (simple, lemma, and word-family counts). However, the current study differed from Treffers-Daller et al.'s (2018) study in three ways.

First, similar to the experiment in chapter 3 (written lexical diversity's role in predicting writing proficiency), this study (chapter 6) investigated spoken lexical diversity and its most directly related language skill (speaking). Second, the study validated the LD measure predictions of speaking proficiency based on three analysis units (simple, flemma, and lemma counts) which only gauge learners' inflectional knowledge. Third, unlike Treffers-Daller et al. (2018) and the writing experiment in chapter 3, the current study was expanded by addressing another important factor (text length effects) that affects LD measures. The reason behind the addition of this factor was that the initial 200-word-analysis findings indicated that the extent to which LD measures predicted speaking proficiency at that length was not significant. Despite the use of the same LD measures and the same cut-off (200 words) point, the LD measures indicated no efficacy in predicting speaking levels for all three analysis

units. This finding revealed the need to examine the minimum text length at which LD measures could predict speaking proficiency .

First, the current study (chapter 6) supported the findings of Treffers-Daller et al. (2018), and the written data analysis in chapter 3, that LD measures' predictive power was influenced by the analysis units. However, the current study on spoken LD and LD measure prediction of speaking proficiency revealed different findings from the writing study in chapter 3. The writing study examining the extent to which LD measures predict writing proficiency indicated that simple, flemma, and lemma counts had different influences on LD measures' writing proficiency predictability.

However, once I examined the extent to which LD measures predicted speaking proficiency following the same procedures as for the writing data, the simple count was found to be a more impactful analysis unit on LD measures' predictive ability, compared to flemma and lemma counts. The Epsilon squared values indicated the greater impact of the simple count on the extent to which LD measures predicted speaking proficiency, since the simple count could best enhance most LD measures' discrimination between speaking levels for all different text lengths except for the 350-word length. All LD measures except for *HD-D* could discern most of the speaking variances on 350-word texts, based on the lemma count. Among the three analysis units, the flemma count appeared to be the least impactful unit on LD measures in discriminating between speaking levels. Furthermore, Kruskal-Wallis test results illustrated that LD measures became more powerful in discriminating between speaking levels as the text lengths increased since the *p* values were smaller on the lengthier texts.

Second, regarding the influences of text length on LD measures speaking predictions based on the three word counting criteria, regression analyses showed that

no LD measures were significant speaking predictors on the short length texts (200, 250, 300) for all three analysis units. However, the combined models of the measures (i.e., best basic measure with best sophisticated measure) turned to be significant from 350 words onwards. For the simple count, *Types* and *D* for 350 tokens, *TTR* and *HD-D* for 400 tokens, and *TTR* and *D* for 450 tokens were significant once they were combined. For the flemma count, *TTR* and *D* for 350 and 400 tokens, and *Types* and *HD-D* for 450 tokens could significantly discern speaking variances as combined measures. For the lemma count, *TTR* and *HD-D* on 350 and 400 tokens, and *TTR* and *MTLD* on 450 tokens were significant once combined. When the full length texts of different token numbers were analyzed, *TTR* and *HD-D* were each significant both as separate and combined speaking predictors, based on simple and flemma counts. For the lemma count, *TTR* and *D* could significantly predict the speaking levels once analyzed either separately or in combination.

Overall, the simple count had the greatest influence on the extent to which LD measures predicted speaking proficiency, compared to the flemma and lemma counts, both for the short spoken text lengths (200, 250, 300 words) and for the longer spoken samples (400-, 450-word, and full lengths). The lemma count was the most discriminating unit on the extent to which LD measures predict speaking proficiency for 350 tokens. The combination of one basic and one sophisticated LD measure could significantly predict speaking levels once the texts were at least 350 words long. Indeed, the measures were useful both as separate and combined speaking indicators for full length texts of varying token numbers.

6.5 Limitations

This study represents an early attempt to address the influence of the analysis unit and text length on the extent to which LD measures predict speaking proficiency;

however, the study includes at least two limitations that might affect the generalizability and reliability of the findings.

The first limitation relates to the small participant number (N = 55). Because of the limited availability of spoken recordings, the participant sample size was small for each of the IELTS levels: 6.5 (N = 17), 7 (N = 19), and 7.5 (N = 19). To ignore normality assumption violations of the data, I applied non-parametric statistical analyses to the small data sets although they might indicate less statistical power. The study cannot explore L1 background influences on the extent to which LD measures predict speaking proficiency because the participant sample was too small to form different L1 background groups. The current study investigates a small participant group with mixed L1 backgrounds, and the findings may thus seem less reliable and generalizable, so future research should conduct wider studies with larger participant sample sizes and analyses of different L1 background groups.

Second, the study examines only the presentation parts of the spoken recordings, excluding the discussion parts. In reality, the class teacher(s) assigned the speaking proficiency levels, based on their evaluation of both presentation and discussion parts. However, the current study did not consider the number of different words used in the discussion parts. Due to this weakness, some important or useful information might be missing.

6.6 Conclusion

The current partial replication study examined the extent to which LD measures predict speaking proficiency depending on the various analysis units and text lengths used. The study indicated that LD measures' predictive power is indeed influenced by the analysis units and text lengths. The findings revealed the greatest influence of the simple count on the extent to which LD measures predict speaking proficiency,

compared to the flemma and lemma counts. This finding was consistent across all different constant text lengths except for the 350-word texts, on which the lemma count was the most discriminating analysis unit. Another important finding was that the LD measures indicated the stronger speaking predictive power on the longer constant text lengths (starting from 350 words) once the measures were combined. However, LD measures could predict speaking levels on full length texts separately as well as in combination for all three analysis units. Despite being a small study with some limitations, the current study highlights the importance of the careful consideration of the appropriate analysis units and the minimum constant text length in assessing lexical diversity based on the particular language mode, which, in this study, was speaking.

Chapter 7

Discussion

7.1 Overview

I divide this discussion chapter into three sections. The first section, 7.2, summarizes and discusses the major findings of the four experimental studies (chapters 3, 4, 5, and 6), examining the four different factors affecting LD measures' ability to accurately predict L2 English writing and speaking proficiency. These four factors are the analysis unit, L1 background, language proficiency, and text length. The second section, 7.3, presents three general claims relating to the analysis unit choice in L2 lexical diversity assessment, how the accuracy of LD measure predictions of L2 writing and speaking proficiency may differ, and what the minimal constant text lengths required are for LD measures' accurate prediction of L2 writing and speaking proficiency. The third section, 7.4, highlights four major limitations of the current PhD research to mitigate against overgeneralization of findings, conclusions, and claims. The third section also explains some implications of the findings for LD measure evaluation and validation practices and suggests a future direction for further research to enhance the existing LD research knowledge.

7.2 Summary of findings

The current PhD research validated the extent to which LD measures predict both L2 English writing and speaking proficiency. As discussed in section 2.5, there has been an inadequate effort in previous research to address four factors (analysis unit, L1 background, language proficiency, and text length) that can affect the predictive power of LD measures. In response to such an important gap in LD

measure validation, I conducted four experimental studies to answer each study's specific research questions.

The first experimental study (chapter 3) explored the influences of the choice of analysis unit on the extent to which LD measures predict the writing ability of L2 learners from diverse L1 backgrounds. The study attempted to fill some gaps in Treffers-Daller et al.'s (2018) study, such as the lack of examination of written LD and writing skill relationships and overlooking the suitability of the flemma count in L2 LD assessment. The study investigated whether LD measures were discriminative of IELTS-based writing proficiency and compared the simple, flemma, and lemma count influences on LD measures and scores.

The findings showed that word-counting techniques influenced LD measures and scores, and that LD measures' writing predictions depended on the choice of analysis units once the text length was restricted to 200 tokens. Based on the simple count, *Types, TTR, D,* and *HD-D* could predict the highest level (7.5) and two levels below (6.5 and 7) while *Guiraud's Index* could discriminate between the lowest and highest levels. Once the flemma count was used, all three basic measures, *D*, and *HD-D* were predictive of the highest and two lower levels. Based on the lemma count, which was the most discriminating unit, the three basic measures could discriminate between all three top writing levels, whereas *HD-D* appeared to be a more precise predictor than *D* or *MTLD*, which could predict only the broader 6.5 and 7.5 writing levels. Thus, the basic LD measures appeared to be more powerful writing predictors than the sophisticated measures.

Second, I conducted a follow-up study (chapter 4) to address one of the important limitations of chapter 3: L1 background influences on the extent to which LD measures predict writing proficiency. An earlier study (Yu, 2010) showed that LD

measure predictions of writing proficiency varied depending on learners' L1 backgrounds (Philippines and Chinese). As an attempt to control for such L1 background effects on LD measures, chapter 4 investigated the extent to which LD measures predict writing proficiency of L1 Chinese L2 English learners, under controlled text length based on different analysis units, following the similar procedures used in chapter 3's study.

The findings showed that LD measures seemed better discriminators of the language proficiency of L1 Chinese L2 English writers once flemmatization and lemmatization were applied. Once the data were flemmatized, *Types, D,* and *HD-D* were the most significant writing discriminators, whereas *TTR, Guiraud's Index*, and *MTLD* could best distinguish writing variances based on lemma count. *Types, TTR*, and *Guiraud's Index* were predictive of the highest and lowest levels for both flemma and lemma counts, and *D* could predict two adjacent (7 and 7.5) writing levels. Among the six LD measures, *HD-D* was the best writing indicator because of its fine discrimination between the highest and two lower levels for both simple and flemma counts, whereas *MTLD* predicted no writing levels when applying any of the three analysis units. I found that the LD measures showed a lower writing predictive power for writing proficiency of L1 Chinese L2 English writers compared to that of L2 English writers from multiple L1 backgrounds (see chapter 3).

The third experimental study (chapter 5) validated the extent to which LD measures predict writing proficiency by incorporating all four influential factors in a single study. The study examined the writing proficiency influence on the extent to which LD measures predict writing proficiency of L1 Chinese L2 English writers under the controlled text length condition for three analysis units. Initially, the study investigated how similar or different the variability within the three writing

proficiency levels (6.5, 7, 7.5) was, and then examined the variation in the extent to which LD measures predict writing proficiency for the simple, flemma, and lemma counts.

The findings highlighted that all three writing levels had low variability and that the overall variability between the three writing levels was not significantly different. Regarding LD measures' discrimination between intragroup writing differences, the three analysis units had different influences on different LD measures' capacity to discriminate between IELTS 6.5 low and high groups. *Types* and *Guiraud's Index* were stronger discriminators on flemmatized data, *TTR* and *MTLD* were more powerful on non-lemmatized data, and *D* and *HD-D* could explain more variances of 6.5 writing sub levels. However, five out of the six LD measures could best estimate IELTS 7 within-level (low and high groups) writing differences on the raw data when a simple count was applied, and *HD-D* was the most discriminative of writing sublevels on the lemmatized data. For writing 7.5 level subgroups, the lemma count could best increase all three basic measures' discriminative power of intragroup writing variability, the simple count was more impactful on *TTR* and *MTLD*, and flemma count could best enhance *D* and *HD-D* discrimination between writing sub levels. However, none of the six LD measures could significantly predict low and high groups of all three writing levels.

The fourth study responded to the LD research gaps of the extent to which LD measures predict L2 speaking proficiency based on different analysis units and of the minimum constant spoken text length required for LD measures to predict speaking proficiency. To enhance the comparability of the extent to which LD measures predicted, I sought findings for two separate skills (writing and speaking) based on simple, flemma, and lemma counts, and the study followed similar procedures to the

writing study (chapter 3). The study investigated whether LD measures could predict L2 speaking proficiency by considering not only the same constant text length (200 words) as writing in chapter 3 but also by trialing different constant text lengths and full length with varying tokens.

The findings revealed that the extent to which LD measures predict speaking proficiency was influenced by the analysis unit and text length. Based on the simple count, all three basic measures (*Types, TTR, Guiraud's Index*) and *D* were more discriminative of speaking levels (6.5, 7, 7.5) for most text lengths, except for the 350-word length on which the lemma count was the most distinctive unit. The LD measures were not significant speaking predictors for any of the constant text lengths, and only *TTR* could significantly predict speaking levels once I examined the varying length texts. However, LD measures tended to be more significant predictors once the texts were longer. In particular, the basic LD measures appeared stronger in predicting speaking levels from text lengths of 400-word onwards. Despite the LD measures under investigation being insignificant as L2 speaking proficiency predictors when deployed discretely, regression analysis findings indicated that the LD measures, which were most strongly correlated with speaking scores, achieved stronger and more significant speaking predictions once these measures were combined (see Tables 6.17, 6.18, 6.19).

7.3 General claims

Based on the four empirical studies' findings, I make three general claims that might contribute to the current LD literature. I should also note that further confirmation from wider studies using sufficient participant sample sizes is still needed to make these claims firm. I discuss the following three claims in the separate sub-sections below.

1. One analysis unit might not always be the best unit in L2 lexical diversity assessment.
2. LD measures might be stronger indicators of writing proficiency than speaking proficiency.
3. LD measures require longer constant text lengths for speaking predictions than for writing predictions.

7.3.1 One analysis unit might not always be the best unit in L2 lexical diversity assessment.

I examined LD measures' predictions of writing and speaking proficiency based on a simple count and two alternative units which demand learner inflectional knowledge (flemma and lemma counts). The findings of the four empirical validation studies (chapters 3, 4, 5, 6) support the existing evidence for the importance of word counting criteria in LD assessment in L2 contexts (Treffers-Daller, 2013; Treffers-Daller et al., 2018). Being the first study that fills the important gap of determining the flemma count's usability, the findings add new information to the existing LD body of knowledge about the flemma count's effects on LD measures' capacity to predict L2 writing and speaking proficiency. These useful insights could contribute to the recent ongoing debate regarding analysis unit choice in L2 vocabulary knowledge assessment.

Based on the four empirical studies' findings, just one analysis unit might not always best fit L2 learners who manifest diverse learner characteristics (e.g., L1 background, language ability). More precisely, different analysis units should be adopted appropriately in evaluating the L2 vocabulary range under three different conditions, namely language modes (writing and speaking), L1 background (mixed

versus single L1 background), and language (writing) proficiency levels 6.5, 7, and 7.5 (see Table 7.1).

First, simple, flemma, and lemma counts appeared to have different influences on the extent to which LD measures predict the two distinct skills (writing and speaking). As mentioned above, for greater comparability of findings, I performed similar procedures in both writing and speaking data analyses. Despite having different participants from various L1 backgrounds, I analyzed the same proficiency levels for the two separate skills (IELTS 6.5, 7, and 7.5 bands for both writing and speaking). Both writing and speaking participants responded to the same theme, *"Globalization impact on the society"*, and their written and spoken texts were rated using IELTS-based writing and speaking rubrics. The data cleaning for LD assessment, lemmatization process, and LD score calculation was the same.

However, the simple, flemma, and lemma counts showed different effects on the extent to which LD measures predict writing and speaking proficiency. The lemma count, which could best increase five out of the six LD measures' (*Types, TTR, Guiraud's Index, D, MLTLD*) writing discrimination, appeared to be the most impactful analysis unit. This finding was consistent with Treffers-Daller et al. (2018), which examined written LD and claimed the lemma count as the most discriminating unit in predicting CEFR general proficiency (B1 to C2) levels. However, for speaking, the simple count had the most significant influences on all three basic measures and *D,* whereas *MTLD* and *HD-D* were more discriminative of speaking ability based on the lemma count.

Interestingly, the flemma count, which I had expected to be more suitable for the participants in my studies (IELTS 6.5, 7, 7.5 writing and speaking bands, equivalent to CEFR upper-intermediate B2 and advanced C1 levels), seemed the least

helpful unit. Ishii et al. (2021) highlighted the complex meaning relationships of a word under two different word classes for their low-proficiency L2 learners. The authors argued that a flemma count might not be appropriate for their Japanese university students whose proficiency was equivalent to CEFR A1 and A2 levels. In addition, I found that the flemma count also appeared to be a less effective unit for the more proficient L2 learners in my studies who might have larger lexical knowledge, especially of inflections. The potential reason might be the fact that my participants used only a few flemmas (inflected words in two different parts of speech) in their written as well as spoken outputs, as I noticed during the data cleaning process. However, they repeatedly used some flemmas such as "*developed*" and "*developing*" in different forms (adjective, present, and past participles). It seems to imply that highly proficient L2 learners might use flemmas only when necessary.

Once I controlled for L1 background and analyzed LD measures' discrimination of the writing proficiency of L1 Chinese L2 English learners, flemma and lemma counts indicated greater influences on LD measures' discrimination between the intergroup (6.5 vs 7 vs 7.5) writing differences. Similar to the mixed L1 background group analysis, I found the simple count to be the least useful unit in discriminating between L1 Chinese L2 English writers. This finding implies that a simple count of different words used in the writing samples might not be as helpful as flemma and lemma counts in explaining writing variances.

The investigation of the LD measures' intragroup (e.g., low vs high 6.5 level Chinese writers) writing discrimination revealed mixed findings of the analysis units' suitability for different (6.5, 7, 7.5) levels. All three analysis units had different influences on the LD measures for the lowest writing level of 6.5. The simple count showed greater effects on *TTR* and *MTLD*; the flemma count on *Types* and *Guiraud's*

Index, and the lemma count on *D* and *HD-D*. In contrast, the simple count was the most discriminating unit in differentiating between writing level 7's low and high sublevels, whereas the lemma count appeared to be a more impactful unit for writing level 7.5 sublevel discrimination.

Taken together, my empirical studies' findings highlight it seems inappropriate to assume that one analysis unit is better than other units for L2 English learners in LD assessment. Here, one size does not fit all; a "horses-for-courses" approach would be better. The analysis unit selection is likely to depend on language modes as well as L2 learner characteristics, such as first language background and language proficiency. The findings support Webb (2021), who claimed that the lexical unit choice should depend on several factors, including learner morphological knowledge and proficiency.

Webb has suggested using smaller lexical units for less proficient learners and larger units for more proficient learners. Brown et al. (2021) also claimed the simple count's suitability for young and beginner learners. However, my research revealed some contrasting evidence in that a simple count, which is the smallest word-counting unit, is more effective in discriminating between the speaking proficiency of upper-intermediate and advanced-level L2 English speakers. This implies that a simple count, the lowest level of Lauer and Nation's (1993) scheme, might not always be the most appropriate unit for lower-level English learners. Similarly, counter-intuitively, it appears erroneous to assume that a flemma count, which demands higher inflectional knowledge than a lemma count, is a better unit for higher-level L2 learners.

Overall, LD researchers should not assert whether a simple, flemma, or lemma count is the more suitable unit for evaluating L2 English learners, solely depending on

one factor (e.g., learner word knowledge or language proficiency). Since these analysis units' suitability might vary widely under different circumstances (language modes, L1 backgrounds, and proficiency), researchers should be cautious about making blanket claims about which lexical unit is the best in L2 lexical diversity assessment.

Table 7.1

Most Impactful Analysis Units on LD Measures' Writing and Speaking Discrimination Under Controlled Text Length (200 words)

LD measure	Most discriminating analysis units (Based on the effect sizes)					
	Mixed L1 Speaking	Mixed L1 Writing	L1 Chinese Writing			
			Inter-group (6.5 vs 7 vs 7.5)	Intra-group (Low vs high for 6.5, 7, 7.5)		
				6.5	7	7.5
Types	Simple	Lemma	Flemma	Flemma	Simple	Lemma
TTR	Simple	Lemma	Lemma	Simple	Simple	Simple/ Lemma
Guiraud	Simple	Lemma	Lemma	Flemma	Simple	Lemma
D	Simple	Lemma	Flemma	Lemma	Simple	Flemma
MTLD	Lemma	Lemma	Lemma	Simple	Simple	Simple
HD-D	Lemma	Flemma	Flemma	Lemma	Lemma	Flemma

7.3.2 LD measures might be stronger indicators of writing proficiency than speaking proficiency.

The current PhD research findings show that, based on 200-word text length, the LD measures' capture of the same construct (lexical diversity) showed different predictive levels of L2 proficiency in two distinct language modes (writing and speaking). Neither basic measures (*Types, TTR, Guiraud's Index*) nor sophisticated

measures (*D, MTLD, HD-D*) could significantly predict L2 speaking proficiency. The LD measures tended to be more predictive of the writing proficiency of L2 English learners from both mixed L1 backgrounds and from a single L1 (Chinese) background (see Table 7.2). Regression analyses with LD measures, which were most strongly correlated with writing and speaking scores, confirmed that LD measures could estimate more writing variances than speaking variances (see Table 7.3). Based on these findings, the extent to which LD measures predicted writing and speaking were different. The LD measures might be more powerful L2 writing predictors than speaking predictors when using a constant written and spoken text length of 200 words.

This finding of the LD measures' different predictive capabilities of L2 writing and speaking seems reasonable because these two different modalities each have their own underlying differing characteristics, such as formality and syntactic or lexical features (Drieman, 1962; O'Donnell, 1974). O'Donnell (1974) highlighted the syntactic differences between written and oral products (e.g., the higher incidence of gerunds, participles, and attributive adjectives in written texts, resulting in longer written clauses). A significant earlier study by Drieman (1962) explored written-spoken language distinctions under almost identical conditions and found higher lexical diversification in written texts than spoken texts. This finding provides some support for my finding of LD measures' stronger ability to explain writing differences than speaking.

Regarding the L2 lexical knowledge underpinning these two language modes, a very recent study by Uchihara and Clenton (2022) found no significant differences between the written and spoken productive vocabulary knowledge of L2 university students from various L1 backgrounds. However, their study's visual analysis of the

correlations of these two vocabulary knowledge domains revealed that some learners' written and spoken productive vocabulary knowledge was not identical. Yu (2010) also suggested that the diverse vocabularies of the same group of L2 English learners in written and spoken modes were similar. However, his study revealed one LD measure (*D)* appeared better at predicting speaking proficiency than writing proficiency. Despite support for my claim on LD measures' differing ability to predict writing and speaking, Yu's finding was opposite from my experimental studies' findings that LD measures are stronger predictors of writing proficiency than speaking proficiency.

However, I consider that the findings of Yu's study and my spoken study are incomparable rather than contradictory. There are three potential reasons these findings may not represent a like-for-like comparison. The first reason might be the differences in the spoken test; Yu's finding was based on the analysis of learners' speech in unplanned interviews, whereas my spoken data analysis examined more planned, formal, and academic seminar presentations. The spontaneous speakers might well have needed fixed expressions, which might have resulted in a greater number of different words. However, most presenters in my spoken study might have benefitted from the planned presentation task. They seemed to use fewer fixed expressions, and they could use as many different words as necessary. That might result in the finding that LD measures are comparatively poor at predicting speaking proficiency under the controlled text length in my study.

The second reason might relate to the text length. Yu's finding was drawn from the examination of varying spoken text lengths while Yu restricted the text length to 200 words in my studies. The third reason might be the spoken study's drawback of having compared the extent to which LD measures predict the L2 writing

and speaking proficiency of different participants albeit under almost identical conditions (e.g., the writing and speaking prompt, IELTS levels such as 6.5, 7, 7.5, LD measures, and procedures were all the same).

I can claim that LD measures could better explain writing differences than speaking differences once the tasks are more planned and formal, and once the text length (200 words) is consistent.

Table 7.2

Exploring the Extent to which LD measures Predict Different Writing and Speaking Proficiencies Under Controlled Text Length (200 words)

Measures	Mixed L1 Speaking			Mixed L1 Writing			L1 Chinese Writing		
	6.5-7	6.5-7.5	7-7.5	6.5-7	6.5-7.5	7-7.5	6.5-7	6.5-7.5	7-7.5
Types0	NS	NS	NS	NS	*	*	NS	NS	NS
Types1	NS	NS	NS	NS	*	*	NS	*	NS
Types2	NS	NS	NS	*	*	*	NS	*	NS
TTR0	NS	NS	NS	NS	*	*	NS	NS	NS
TTR1	NS	NS	NS	NS	*	*	NS	*	NS
TTR2	NS	NS	NS	*	*	*	NS	*	NS
Guiraud0	NS	NS	NS	NS	*	NS	NS	NS	NS
Guiraud1	NS	NS	NS	NS	*	*	NS	*	NS
Guiraud2	NS	NS	NS	*	*	*	NS	*	NS
D0	NS	NS	NS	NS	*	*	NS	NS	NS
D1	NS	NS	NS	NS	*	*	NS	NS	*
D2	NS	NS	NS	NS	*	NS	NS	NS	NS
MTLD0	NS	NS	NS	NS	NS	NS	NS	NS	NS
MTLD1	NS	NS	NS	NS	NS	NS	NS	NS	NS
MTLD2	NS	NS	NS	NS	*	NS	NS	NS	NS
HDD0	NS	NS	NS	NS	*	*	NS	*	*
HDD1	NS	NS	NS	NS	*	*	NS	*	*
HDD2	NS	NS	NS	NS	*	*	NS	NS	NS

Table 7.3

Regression Analysis Findings Showing LD Measure Predictions of Writing and Speaking Proficiency (200-Word Length)

Units	Measure	Mixed L1 Speaking	Mixed L1Writing	L1 Chinese Writing
Simple	Single	Types (0.8%) HD-D (0%)	Types* (5.9%) HD-D* (7.4%)	Types (3.9%) HD-D* (5.9%)
Flemma	Single	TTR (0.2%) D (0.2%)	Types* (9.9%) HD-D* (8.4%)	Types* (8.7%) HD-D* (7.2%)
Lemma	Single	TTR (0.1%) MTLD (0%)	G* (12.1%) D* (7.9%)	G* (9.1%) D* (5.3%)

Note. Asterisk (*) shows that LD measures were significant writing or speaking predictors.

7.3.3 LD measures might require longer constant text length to predict speaking proficiency compared to writing proficiency.

Issues about text length variety and LD measures' predictive validity have been ongoing in the LD assessment field. There have been ongoing efforts to overcome the text sample size problem by applying different calculation methods, developing less sensitive LD measures (*D, MTLD, HD-D*), and/or suggesting text length standardization (Treffers-Daller, 2013; Treffers-Daller et al., 2018). Some LD researchers have tried to fine-tune and suggest the written text lengths to which LD measures showed more stability (Koizumi, 2012; Koizumi & In'nami, 2012) and the minimum written text lengths required for several existing LD measures (Zenker & Kyle (2021).

In another study, Treffers-Daller et al. (2018), has considered consistent text length and has shown that LD measures could predict CEFR overall levels based on a short cut-off point (200 words). However, the question remains whether LD measures have similar or different predictive levels of L2 writing and speaking proficiency when using the same constant text lengths. Although this text length issue has long been acknowledged, most LD studies have barely considered text length's effects on the extent to which LD measures predict speaking proficiency, as I highlighted in section 2.5, and so provided LD measure speaking predictions based on varying text lengths. I, therefore, do not know how long a spoken text length should be, at least, in order for LD measures to be effective predictors of L2 speaking proficiency and what the constant text lengths should be for comparison.

I therefore investigated the extent to which LD measures indicated speaking proficiency, evaluating both different constant lengths (from 200 to 450 words) and varying full lengths (see chapter 6). Based on my research findings, the minimum constant text lengths at which LD measures produced effective writing and speaking predictions were not equivalent. When 200-word essays were examined, the LD measures could discriminate between the writing levels of L2 English learners of mixed L1 backgrounds (see chapter 4) as well as L1 Chinese L2 English learners (see chapter 5). However, in the examination of the same constant text length for speaking, the LD measures failed to predict the speaking levels of L2 English speakers from various L1 backgrounds, regardless of the analysis unit used.

Thus, unlike for writing predictions, the 200-word text length seemed insufficient for LD measures to predict L2 speaking proficiency. However, the LD measures' predictive power increased once the spoken texts were longer, particularly for the basic LD measures. The basic LD measures showed a tendency to be

significant from 400 words, but of these, only *TTR* was effective in predicting L2 speaking proficiency for varying text lengths. Regression analysis findings indicated that even the basic and the sophisticated measures, which had the highest correlations with writing and speaking, could not significantly predict L2 speaking levels when deployed as separate measures (see Tables 6.17, 6.18, 6.19). However, these measures appeared useful once combined.

Overall, the LD measures did not predict speaking proficiency at the constant text length (200 tokens) at which the LD measures were predictive of L2 writing. Since the LD measures turned out to be more significant speaking predictors for lengthier texts, longer constant text length appears more necessary for LD measures to predict L2 speaking proficiency than for L2 writing proficiency.

7.4 Limitations, implications, and recommendations

To reduce the overgeneralization of the current PhD research findings, conclusions, and claims, I have acknowledged each experimental study's limitations in sections 3.6, 4.5, 5.5, and 6.5. In addition, in this current section, 7.4, I highlight four major limitations which might have influenced my studies' findings.

First, the LD measures' intra-group writing variability (chapter 5) and speaking (chapter 6) predictions were examined using small participant sample sizes (N = 103, 55). The participant numbers in each writing subgroup (low or high) of the three writing levels are small and unbalanced, leading to the low within-group writing variations. Similarly, the LD measures' speaking prediction findings are based on small participant sample sizes [6.5 (N = 17), 7 (N = 19), 7.5 (N = 19)]. The findings might be of low generalizability and might be different for larger studies with sufficient participant numbers.

The second limitation relates to the writing and speaking tests, specifically the "pre-writing and pre-speaking planning". With the benefit of that planning time, my participants might have used more diverse vocabulary and more derivations by consulting dictionaries and thesauruses, rather than having to depend on and represent their actual use of their vocabulary range and derived forms as they would have to in doing spontaneous tasks. There has been rich evidence of planning's enhancing effects on L2 writing or speaking performance. Ellis and Yuan (2004) showed that, for L1 Chinese L2 English writers under two planning conditions (pre-task and online planning), the pre-planners outperformed the non-planned writers in terms of fluency, accuracy, and complexity. Similarly, Seyyedi et al. (2013) claimed that writers could benefit from pre-task planning and produce more fluent and complex written texts. Limpo and Alves (2018) also highlighted that writers who had engaged in planning strategies could produce larger word numbers per clause than writers in a no-planning condition. Therefore, the current PhD research findings or conclusions might not be generalizable to unplanned writing or speaking assessments.

The third limitation lies in the spoken data collection process. I analyzed only the presentation parts of the spoken transcripts, excluding the discussion parts, which included more moderator or interviewer speeches and interactions. Conversely, the experienced speaking rater (not the researcher) assessed both presentation and discussion parts and assigned the three speaking (6.5, 7, 7.5) bands that I examined. The investigation of the lexical diversity used in only the presentation part might not represent the participants' speaking proficiency levels, which were evaluated based on both their presentation and discussion skills. That might be a plausible reason behind the reported findings here of LD measure predictions of speaking proficiency not being significant.

The fourth limitation is that the speaking study (chapter 6) could not address L1 background and language (speaking) proficiency influences on LD measures' speaking predictions. Initially, I had intended to explore LD measures' speaking predictions while addressing the four influential factors (analysis unit, L1 background, speaking proficiency, and text length). Since the data set is small because of data availability and the intentional use of the same writing and speaking levels (IELTS 6.5, 7, 7.5) to establish compatibility and comparability, it was impossible to arrange and analyze different L1 groups and speaking sub levels. Therefore, the spoken data analysis does not provide information on L2 English learners' L1 background impact on LD measure predictions of speaking proficiency.

Despite these limitations, I believe this PhD research provides some important implications for guiding future LD measure validation and LD assessment in L2 contexts.

First, my partial replication studies of a prior research study by Treffers-Daller et al. (2018) confirm their claim that lemmatization techniques influenced LD measure predictions of L2 language proficiency. My studies validate LD measure predictions under different conditions, such as different L2 learners, language skills (writing and speaking), a lemmatization program (Python), and the use of a flemma count. This dependence of LD measures' writing and speaking predictions on the particular lexical units analyzed supports Treffers-Daller et al.'s (2018) finding of LD measures' greater predictions of L2 language (general) proficiency based on lemma counting. Thus, the current PhD research adds empirical evidence of LD measures' applicability in L2 writing and speaking proficiency assessment.

Second, the current research responds to the emerging need to address four influential factors (analysis unit, L1 background, language proficiency, and text

length) in L2 lexical diversity assessment. In the literature section, I have identified the inadequate attempts to limit these four factors in validating LD measures' usefulness in predicting L2 language proficiency. Previous studies' findings of LD measures' predictive validity have been based on studies addressing only one or two of these four factors. That might be a potential reason for the mixed or inconsistent findings and conclusions of these studies of LD measures' L2 language predictions. Therefore, the significance of my research lies in incorporating all four factors in L2 writing assessment. The findings yield new information about the variation of LD measures' predictability across these four different factors (analysis unit, L1 background, language proficiency, and text length). LD researchers should carefully interpret these findings of LD measure predictions that show how they can vary depending on how potential factors and L2 learners' unique characteristics are controlled.

Third, this PhD research contributes to the current controversial discussion on "the best analysis unit choice in L2 vocabulary assessment". The study adds new, nuanced information to the existing LD literature about how the flemma count affects LD measures' capacity to predict L2 writing and speaking proficiency. Many previous LD studies paid little attention to the criticality of the choice of the lexical unit to be analyzed in capturing L2 learners' existing word knowledge. However, recent research evidence of L2 learners' challenges with derivations (McLean, 2017; Schmitt & Zimmerman, 2002; Stoeckel et al., 2020) has implied that the analysis unit used should match L2 learners' inflectional and derivational knowledge. For instance, a word-family count might not be as discriminative as the smaller units (e.g., simple, flemma, or lemma counts) for low proficiency learners with limited derivational knowledge.

This issue of the analysis unit selection has also gained an interest in the LD research field. Treffers-Daller et al. (2018) explored different analysis units' suitability in LD assessment and suggested the use of a lemma count over simple and word-family counts for CEFR B1 to C2 level learners. However, no single LD study has examined the influences of a flemma count on LD measures' ability to predict L2 writing and speaking. The current research is the first study to explore the flemma count. This study found that a flemma count, which requires learner word class knowledge of inflections, seems to be the least useful in predicting the writing and speaking proficiency of L2 learners from mixed L1 backgrounds. However, the flemma count appears to be more effective in discriminating between intergroup writing levels of L1 Chinese L2 English learners. These findings imply that the flemma count's suitability is likely to be influenced by learners' L1 background. Therefore, we should not always assume that bigger analysis units are better for L2 learners with higher inflectional knowledge.

Fourth, the current PhD research raises one important question that might be helpful for guiding future LD assessment and measure development. First, the reported negative correlations between most LD measures and writing and speaking leave an intriguing question to be resolved: Should we consider not only lexical diversity quantity (number of different words) but also lexical diversity quality (e.g., how frequent, advanced, or concise different words are) in LD evaluation?

For both mixed and single L1 backgrounds, I found that IELTS 7.5 level essays had the lowest lexical diversification compared to the two lower level (6.5 and 7) essays when evaluated by most LD measures and by all three word counting units. For spoken transcripts, the highest speaking level (7.5) received the lowest *Types, TTR*, and *Guiraud's Index* scores for all text lengths. These findings contradict

Treffers-Daller et al. (2018), which reported the more diverse vocabulary use of higher-proficiency L2 learners compared to lower-proficiency learners. More advanced L2 learners might well be assumed to have larger vocabulary sizes than low-proficiency learners. If it is the case, the expectation would be that more advanced (7.5 level) writers or speakers in my study should have produced a greater number of different words in their written or spoken products than the 6.5 and 7 level writers or speakers; however, this expectation was confounded.

There are two plausible reasons that might help explain this apparent contradiction. First, the aspect of vocabulary is just one of the assessment criteria for writing or speaking, so vocabulary alone might not fully manifest writing or speaking proficiency. 7.5-level essays might have gained more points on other evaluation criteria (e.g., task fulfillment). The second potential reason might relate to participants' vocabulary use in writing or speaking. Lower-proficiency learners' writing might have been wordy because of the overuse of prepositional phrases, lengthier modifiers, or unnecessary words or phrases. The higher-proficiency learners could probably produce more concise writing to deliver their intended message effectively. For instance, advanced writers might use the shorter phrase, "*Proficiency level classification is important*…" instead of the longer phrase, "*It is important to classify proficiency level*…" used by lower proficient writers. In such a case, simply counting the number of different words might not fully predict writing or speaking proficiency.

Therefore, to contribute usefully to the existing LD research knowledge, it might worth exploring both lexical diversity quality and quantity aspects in estimating L2 writing or speaking proficiency. High-proficiency L2 learners might have greater productive vocabulary knowledge and so might be capable of using low frequent,

academic, advanced, or concise words to convey their meaning effectively compared to low-proficiency learners. That might reduce the total number of different words used in their writing or speaking products, resulting in this lower lexical diversification of their texts. Therefore, totting up the numbers of different words might not be enough to accurately evaluate or estimate writing or speaking proficiency. Thus, future studies should examine the aspects of lexical diversity's quality in predicting L2 language proficiency as well as that of lexical diversity's quantity.

Furthermore, LD researchers have long reported the potential lexical diversity knowledge differences between L2 learners from various L1 backgrounds. Yu (2010) compared two different L1 groups (Filipinos and Chinese) and found no significant differences in their written lexical diversity (*D*) scores despite significant LD score differences for all learner groups comprising mixed L1 backgrounds. My studies revealed that LD measures had a lower capacity to discriminate between L1 Chinese L2 English writers than L2 English learners from diverse L1 backgrounds. Further studies should analyze how the similar or different lexical diversity knowledge of L2 learners from different L1 backgrounds could impact the efficacy of LD measures and measurement. More studies on different L1 background influences on LD scores and measures are required to provide insights on LD measures' applicability to other L1 backgrounds, such as Japanese, Thai, and South Korean.

Future research should fill an important gap left by my speaking study: the lack of research considering the influences of L1 background and speaking proficiency on LD measures' speaking predictions based on different word-counting techniques. Although I had initially aimed at addressing all four factors in examining the LD measure predictions of L2 speaking proficiency, L1 background and speaking

sublevel classification turned out to be unrealizable variables of the study due to the small participant sample size. Therefore, future research should investigate whether LD measure predictions of speaking proficiency is influenced by L1 background and speaking proficiency levels.

Chapter 8

Conclusion

This PhD dissertation has attempted to fulfill the emerging research need of addressing the variability of four influential factors on LD measurement: analysis unit, L1 background, language proficiency, and text length. The experiments in the dissertation were based on Treffers-Daller et al. (2018), which has contributed significantly to L2 lexical diversity research with the insights on how analysis unit choice can affect LD measure predictions of overall L2 proficiency. In particular, the dissertation has aimed at investigating the influences of all four factors that can variously affect LD measures' ability to predict L2 writing, and it has addressed analysis unit and text length influences on LD measures' ability to predict L2 speaking.

To achieve these aims, I first reviewed the existing LD literature and identified the extent to which several scholarly papers controlled these four factors in examining various LD measures' predictive validity. The review showed that L1 background and text length issues have long been acknowledged in LD research field; however, the attempts to control these two factors have been inadequate. Except for Treffers-Daller et al. (2018), the reviewed LD studies did not consider the text length sensitivity of LD measures.

The review also highlighted that different word (type) counting methods have become a pressing recent concern, since LD measure predictive capability seems dependent on how we count the different words used in a written or spoken text. Recent evidence of L2 learners' limited derivational knowledge (McLean, 2017; Ward & Chuenjundaeng, 2009; Schmitt & Zimmerman, 2002; Sukying, 2018) suggests the necessity of applying a lexical unit that captures L2 learners' existing

vocabulary (inflections and derivations) knowledge. Most of the reviewed studies employed a simple type count for L2 English writers or speakers at different proficiency levels although they must have varying knowledge and usage levels of affixation.

Treffers-Daller et al. (2018) conducted a wider investigation by using three analysis units (simple, lemma, and word-family counts) and indicated the different influences of these three units on LD measures, highlighting the greater suitability of a lemma count for high-proficiency L2 learners. In addition, the review clearly showed that what appears to be missing in LD assessment is the exploration of whether a flemma count might better suit L2 learners if a lemma count might underestimate, or a word-family count might overestimate, L2 learner lexical knowledge.

The reviewed LD studies put more emphasis on LD measures' discrimination between different proficiency levels than on LD measures' usability in predicting intragroup variations. Due to L2 learners' wide and diverse proficiency levels, it is also necessary to know which particular LD measure and analysis unit should be adopted to estimate within-level differences.

The review concluded that previous LD studies considered only one or two of the four factors despite previous studies' findings of how useful LD measures can be in predicting L2 language proficiency. There is a considerable need to investigate the flemma count's impact on LD measure predictions of both L2 writing and speaking. I attempted to address all four equally important factors, which had not been systematically considered by any of the existing LD research. Four experimental studies were conducted to investigate six different LD measures predictions: (i) of

writing and speaking proficiency of L2 English learners from multi-L1 backgrounds, and (ii) of inter- and intra-group writing variability of L1 Chinese L2 English learners.

The first experiment, in chapter 3, addressed the analysis unit and text length effects by exploring LD measure predictions of the writing ability of L2 English learners from mixed L1 backgrounds based on simple, flemma, and lemma counts under controlled text length. The findings indicated lemmatization could best enhance LD measures' discrimination and prediction of writing ability compared to non-lemmatization and flemmatization rules. Based on simple counts, *Types, TTR, D*, and *HD-D* could predict IELTS 7.5 level and two lower (6.5 and 7) levels while Guiraud's Index could discriminate between the highest (7.5) and the lowest (6.5) writing levels. Once a flemma count was used, all LD measures except *MTLD* were discriminative of 7.5 and two lower levels. All three basic measures could explain the variances between all three writing levels, and *HD-D* was predictive of 7.5 and two lower levels when data were lemmatized. However, *MTLD* was the least effective, being predictive of only 6.5 and 7.5 levels.

This finding of the lemma count exerting the most significant influence on LD measure predictions of writing proficiency is in line with Treffers-Daller et al. (2018). This finding rather confounds the expectation that a flemma count, which demands slightly higher inflectional knowledge, might be a more discriminating unit for proficient L2 English learners. Regression analyses indicated that *HD-D* was a better writing predictor than *Types* on non-lemmatized data, *Types* on flemmatized data and *Guiraud's Index* on lemmatized data were stronger writing indicators than *HD-D* and *D*. Based on the findings, I have therefore concluded that a lemma count appears to be a more impactful unit on most LD measure predictions of writing proficiency of L2 learners from various L1 backgrounds.

These findings were based on the analysis of data from L2 learners from diverse L1 backgrounds, so it might not be generalizable or applicable to L2 learners from a single L1 background because of the potential L1 background influence on LD measures (Yu, 2010). I therefore extended the study by controlling one more factor (L1 background) in the second experiment (chapter 4). Chapter four restricted the L1 background to only Chinese by focusing on L1 Chinese L2 English writers. The study investigated the diverse influences of simple, flemma, and lemma counts on LD measure predictions of writing proficiency of L1 Chinese learners, using consistent text sample size.

The findings indicated that both flemmatization and lemmatization influenced LD measures, but they exert different influences on different LD measures. Flemma counts could best increase the writing predictive ability of *Types, D*, and *HD-D,* whereas *TTR, Guiraud's Index*, and *MTLD* could best estimate writing differences based on lemma counts. Compared to the preceding multi- L1 background group analysis, LD measures seemed less discriminative of L2 writing ability for L1 Chinese learners. Based on simple counts, only *HD-D* could finely discriminate between 6.5 and 7.5, and 7 and 7.5 levels. Once a flemma count was applied, *Types, TTR*, and *Guiraud's Index* could differentiate between the lowest and highest levels, and *D* could predict two adjacent levels (7 and 7.5). Among all six LD measures, *HD-D* could explain writing differences between 7.5 and two lower levels. Based on the regression analysis findings, *HD-D* was a better writing predictor on non-lemmatized data, *Types* could explain more writing variation on flemmatized data, and *Guiraud's Index* was more discriminative of writing on lemmatized data.

The findings of these LD measures' low predictions of L1 Chinese L2 English writing confirmed the LD measures' dependence on L1 backgrounds. The LD

measures' greater writing predictions of a wider population (L2 learners from mixed-L1 backgrounds) might be because of those learners being of different L1 background groups (e.g., Taiwanese or Thai). The lemma count was found to be the most appropriate unit for the multi-L1 background group while the flemma count appeared to be a more discriminating unit for L1 Chinese L2 English learners.

The experimental study in chapter 5 progressed to incorporate all four influential factors on LD measure predictions of writing proficiency. The study investigated whether LD measures were reliable predictors of intragroup (e.g., between low and high 6.5 writing sublevels) writing variations of L1 Chinese L2 English learners. Additionally, the study investigated whether LD measures predict within-group writing proficiency depending on three different analysis units while text length was controlled.

The findings revealed that the three analysis units' influences on LD measures' discrimination of within-group writing variations were different depending the L2 learners' writing proficiency levels. The simple count-based analysis showed the mixed findings of the three analysis units' effects on LD measures' accuracy (simple count suitability for *TTR* and *MTLD*, flemma count for *Types* and *Guiraud's Index*, and lemma count for *D* and *HD-D*). However, five out of the six LD measures could better discriminate between IELTS writing 7 sublevels based on a lemma count. Moreover, the lemma count could enhance all three basic measures' ability to discriminate between low and high 7.5 writing sublevels, whereas simple and flemma counts could not.

However, the LD measures could not significantly predict the intragroup writing variations of all three levels. This finding was consistent for all three analysis units. Nevertheless, *Types* and *HD-D*, once combined, were significant predictors of

low and high IELTS writing 7.5 sublevels. Based on these findings, I have concluded that LD measure predictions of intragroup writing proficiency and analysis unit suitability depend on language (writing) proficiency.

The last experimental study, in chapter 6, was the first study to date to address the three analysis units' influences on the extent to which LD measures predict L2 speaking proficiency. The study's procedures (IELTS-based proficiency levels, data cleaning, lemmatizations, LD measures) were similar to the LD measure writing prediction experiment reported in chapter 3, except for studying different participants so as to compare LD measures' writing and speaking predictions. Because of limited participant numbers, L1 background and language (speaking) proficiency effects on LD scores and LD measures could not be addressed. However, the study did examine the minimum constant text length that LD measures require to predict speaking proficiency. It was because, unlike writing, the 200-word constant text length was insufficient for LD measures to show the ability to predict speaking.

The findings indicated that a simple count appeared to be the most discriminating unit on all text lengths except the 350-word length, on which a lemma count was the most impactful unit. Using a simple count could make *Types, TTR, Guiraud's Index*, and *D* more powerful in discriminating speaking, whereas a flemma count had more significant influences on *MTLD* and *HD-D*. Regarding LD measure speaking predictions, only *TTR* could significantly predict speaking levels (6.5 and 7.5) for all three analysis units once the spoken text sample size was not controlled. However, basic LD measures became more significant (smaller p-values) on longer constant text lengths from 400 words upwards based on simple and lemma counts and on 450-word length for a flemma count.

Comparing LD measures' writing (chapter 3) and speaking predictions, the lemma count seemed more useful in predicting writing while the simple count was more discriminating of speaking. Despite LD measures' significant writing predictions, the measures failed to predict speaking levels based on 200-word text length analysis. I have therefore concluded that the LD measures have different predictive powers for writing and speaking proficiency, and that the 200-word constant spoken text length was insufficient for achieving LD measure speaking predictions.

Overall, the four experimental chapters' findings highlighted that LD measure predictions of L2 language proficiency were influenced by the factors of analysis unit, L1 background, language proficiency, and text length. The lemma count seemed more appropriate to accurately capture the written LD role in L2 writing assessment, whereas the simple count appeared to be more effective in indicating the correlation between spoken LD and L2 speaking proficiency. Furthermore, the minimum constant text lengths at which LD measures showed writing (200 words) and speaking predictions (350 words) were not the same, and LD measures required longer constant text length for speaking predictions than writing predictions.

However, the generalizability of these four experimental studies' findings and conclusions might be limited for some reasons. A larger study with greater participant numbers might yield different findings, particularly for the experiments in chapters 5 and 6. Furthermore, this current PhD study focused on only planned writing and speaking tasks, so the reported findings of analysis unit, L1 background, language proficiency, and text length influences on LD measures' L2 language predictions might not be applicable to the analysis of a spontaneous task. Moreover, the spoken study (chapter 6) examined only different words used in the presentation part without

considering the discussion part in which some presenters spoke less than the moderator or questioner. That might have affected the study's findings. Additionally, the speaking study fails to provide information on how LD measure speaking prediction was influenced by L1 background and speaking proficiency.

Despite these limitations, this PhD study provides a contribution to LD assessment and LD measure validation in the L2 context. First, the study adds empirical evidence to support the importance of the analysis unit choice in L2 lexical diversity assessment, it incorporates four widely acknowledged factors on LD measure predictions of L2 language proficiency, and it provides some previously lacking evidence for the applicability of the flemma count in L2 lexical diversity assessment. Furthermore, the reported paradoxical findings of higher-proficiency L2 writers or speakers using less diverse vocabulary than lower-proficiency learners raises an important question. Future studies should explore whether it is necessary also to consider the quality (frequency, sophistication, or conciseness) of different words deployed in a written or spoken text in estimating L2 language proficiency, instead of counting the numbers of different words. For greater contributions to the LD assessment and validation field, more studies are required to analyze L1 background effects on LD measures by investigating various L1 backgrounds. Future research should address the weakness of my spoken data analysis (chapter 6) by examining L1 background and speaking proficiency influences on LD measures' ability to predict speaking.

Based on the four experimental studies' findings, I conclude that LD measures were useful in predicting L2 language proficiency, especially writing; however, their predictive validity was influenced by the four factors of analysis unit, L1 background, language proficiency, and text length. I also suggest selecting the appropriate analysis

unit depending on these four factors because of the reported diverse findings of the suitability of simple, flemma, and lemma counts. The lemma count was a more impactful unit in discriminating the writing proficiency of L2 learners from multiple L1 backgrounds, whereas the flemma count was a more distinctive unit for L1 Chinese L2 English writers. However, the simple count was a more discriminating unit for L2 English speakers from mixed L1 backgrounds. Moreover, LD measures were found to be stronger writing indicators than speaking indicators when the short stable text length (200 words) was examined.

These four factors cause this variation in the accuracy of LD measures to predict L2 language proficiency. I therefore assert that researchers need to control as many of these factors affecting LD measurement as possible in attempting to validate LD measures' accuracy in predicting L2 language proficiency of L2 learners with such diverse characteristics (e.g., L1 background, language ability, lexical knowledge).

References

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning, 67*(1), 180-208. https://doi.org/10.1111/lang.12232

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical data base (CD-ROM). *Linguistic Data Consortium.* http://hdl.handle.net/21.11116/0000-0001-91EF-E

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

Brown, D. (2018). Examining the word family through word lists. *Vocabulary Learning and Instruction, 7*(1), 51-65. https://doi.org/10.7820/vli.v07.1.brown

Brown, D., Stewart, J., McLean, S., & Stoeckel, T. (2021). The coming paradigm shift in the use of lexical units. *Studies in Second Language Acquisition, 43*(5), 950-953. doi.org/10.1017/S0272263121000668

Brown, D., Stoeckel, T., McLean, S., & Stewart, J. (2022). The most appropriate lexical unit for L2 vocabulary research and pedagogy: A brief review of the evidence. *Applied Linguistics*, *43*(3), 596-602. doi.org/10.1093/applin/amaa061

Clavel-Arroitia, B., & Pennock-Speck, B. (2021). Analysing lexical density, diversity, and sophistication in written and spoken telecollaborative exchanges. *Computer Assisted Language Learning Electronic Journal (CALL-EJ), 22*(3), 230-250. https://www.researchgate.net/publication/354986356

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

Crosssley, S. A., & McNamara, D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing, 18*(2), 119-135. https://doi.org/10.1016/j.jslw.2009.02.002

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2011). Predicting the proficiency level of language learners using lexical indices. *Language Testing, 29*(2), 243-263. https://doi.org/10.1177/0265532211419331

Crossley, S., & McNamara, D. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology, 17*(2), 171-192. http://dx.doi.org/10125/44329

Csomay, E., & Prades, A. (2018). Academic vocabulary in ESL student papers: A corpus-based study. *Journal of English for Academic Purposes*, *33*, 100-118. https://doi.org/10.1016/j.jeap.2018.02.003

Daller, H., & Xue, H. (2007). Lexical richness and the oral proficiency of Chinese EFL students. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 150-164). Cambridge University Press.

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

deBoer, F. (2014). Evaluating the comparability of two measures of lexical diversity. *System, 47*, 139–145. https://doi.org/10.1016/j.system.2014.10.008

Douglas, S. R. (2016). The relationship between lexical frequency profiling measures and rater judgments of spoken and written general English language proficiency on the CELPIP-General test. *TESL Canada Journal, 32*(9), 43-64.

Drieman, G. H. J. (1962). Differences between written and spoken languages: an exploratory study. *Acta Psychologica*, *20*, 36-57.

https://doi.org/10.1016/0001-6918(62)90006-9

Dugast, D. (1978). Sur quoi se fonde la notion d́etendue theóretique du vocabulaire? [What is the basis for the notion of theoretical scope of vocabulary?]. *Francais (Le) Moderne Paris, 46*(1), 25-32.

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

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

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

Gerasimos, F., 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

Gonzalez, M. C. (2017). The contribution of lexical diversity to college-level writing. *TESOL Journal, 8*(4), 899-919. https://doi.org/10.1002/tesj.342

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

Guiraud, P. (1954). *Les Caractères Statistiques du Vocabulaire. Essai de méthodologie* [The statistical characteristics of vocabulary: An essay in methodology]. Presses Universitaires de France.

Ha, H. S. (2019). Lexical richness in EFL undergraduate students' academic writing. *English Teaching, 74*(3), 3-28. DOI: 10.15858/engtea.74.3.201909.3

Herdan, G. (1960). *Type-token mathematics: A textbook of mathematical linguistics*. Mouton.

Higginbotham, G., & Reid, J. (2019). The lexical sophistication of second language learners' academic essays. *Journal of English for Academic Purposes, 37*, 127-140. https://doi.org/10.1016/j.jeap.2018.12.002

Ishii, T., Bennett, P., & Stoeckel, T. (2021). Challenges in the assumptions of using a flemma-based word counting unit. *Vocabulary Learning and Instruction, 10*(1), 1-15. https://doi.org/10.7820/vli.v10.1.Ishii

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. (2013). Defining and measuring lexical diversity. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 13-44). John Benjamins.

Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing, 34*(4), 537-553. https://doi.org/10.1177/02655322/10632

Johnson, W. (1944). Studies in language behavior: A program of research. *Psychological Monographs, 56*(2), 1-15.

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction, 1*(1), 60-69. http://dx.doi.org/10.7820/vli.v01.1.koizumi

Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens, *System, 40*(4), 554-564. https://doi.org/10.1016/j.system.2012.10.012

Kremmel, B., & Schmitt, N. (2016). Interpreting vocabulary test scores: What do various item formats tell us about learners' abilities to employ words? *Language Assessment Quarterly, 13*(4), 377-392. https://doi.org/10.1080/15434303.2016.1237516

Kremmel, B. (2021). Selling the (word) family silver? A response to Webb's "Lemma dilemma". *Studies in Second Language Acquisition, 43*(5), 962-964. https://doi.org/10.1017/S0272263121000693

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*, 757-786. http://dx.doi.org/10.1002/tesq.19

Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing*, *33*(3), 319-340. https://doi.org/10.1177%2F0265532215587391

Kyle, K. (2018). *TAALED: Tool for the analysis of lexical diversity (beta version 1.2.4, Python-based).*

Kyle, K., Crossley, S., & 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

Lai, S. A., & Schwanenflugel, P. J. (2016). Validating the use of D for measuring lexical diversity in low-income kindergarten children. *Language, Speech, and Hearing Services in Schools, 47*(3), 225-235. https://doi.org/10.1044/2016_LSHSS-15-0028

Laufer, B., & Cobb, T. (2019). How much knowledge of derived words is needed for reading? *Applied Linguistics*, 41(6), 971-998. https://doi.org/10.1093/applin/amz051

Laufer, B. (2021). Lemmas, flemmas, word families, and common sense. *Studies in Second Language Acquisition, 43*(5), 965-968. https://doi.org/10.1017/S0272263121000656

Leontjev, D., Huhta, A., & Mantyla, K. (2016). Word derivational knowledge and writing proficiency: How do they link? *System, 59*, 73-89. https://doi.org/10.1016/j.system.2016.03.013

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

MacWhinney, B. (2000). *The CHILDES project volume I* (3rd ed.). Psychology Press.

Malvern, D., & Richards, B. J. (1997). A new measure of lexical diversity. *British Studies in Applied Linguistics*, *12*, 58-71.

Malvern, D., Richards, B. J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development*. Palgrave Macmillan.

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)* [Doctoral dissertation, The University of Memphis].

McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation of vocd. *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., Watanabe, S., & Lamkin, T. A. (2012). The Gramulator: A tool to identify differential linguistic features of correlative text types. In *Applied natural language processing: Identification, investigation, and resolution* (pp. 312-333). IGI Global.

McCarthy, P. M., & Jarvis, S. (2013). From intrinsic to extrinsic issues of lexical diversity assessment: An ecological validation study. In S. Jarvis and M. Daller (Eds), *Vocabulary knowledge: Human ratings and automated measures* (pp. 45-77). John Benjamins.

McLean, S. (2017). 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

Meara, P., & Miralpeix, I. (2016). *Tools for researching vocabulary*. Multilingual Matters.

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*, 1-13. https://doi.org/10.1016/j.asw.2022.100640

Nasseri, M., & Thompson, P. (2021). Lexical density and diversity in dissertation abstracts: Revisiting English L1 vs. L2 text differences. *Assessing Writing, 47*, 1-11. https://doi.org/10.1016/j.asw.2020100511

Nation, P. (2006). *BNC-based word lists*. Wellington: Victoria University of Wellington**.**

Nation, P. (2021). Thoughts on word families. *Studies in Second Language Acquisition, 43*(5), 969-972. https://doi.org/10.1017/S027226312100067X

O'Donnell, R. C. (1974). Syntactic differences between speech and writing. *American Speech*, *49*(1/2), 102-110. https://www.jstor.org/stable/3087922

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

Read, J., & Nation, P. (2006). An investigation of the lexical dimension of the IELTS speaking test. *IELTS Research Reports, 6*, 207-231.

Reid, J. (1986). Using the writer's workbench in composition teaching and testing. In C. W. Stansfield (Ed), *Technology and language testing* (pp. 167-188).

Richards, B. (1987). Type/token ratios: What do they really tell us? *Journal of Child Language, 14*(2), 201-209. https://doi.org/10.1017/S0305000900012885

Sasao, Y., & Webb, S. (2017). The word part levels test. *Language Teaching Research, 21*(1), 12-30. https://doi.org/10.1177/1362168815586083

Schmitt, N., & Zimmerman, C. B. (2002). Derivative word forms: What do learners know? *TESOL QUARTERLY, 36*(2), 145-171. https://doi.org/10.2307/3588328

Seyyedi, K., Ismail, S. A. M. M., Orang, M., & Nejad, M. S. (2013). The effect of pre-task planning time on L2 learners' narrative writing performance. *English Language Teaching, 6*(12), 1-10. http://dx.doi.org/10.5539/elt.v6n12p1

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

Sukying, A. (2018). The acquisition of English affix knowledge in L2 learners. *NIDA Journal of Language and Communication, 23*(34), 89-102.

Tavakoli, P., & Uchihara, T. (2020). To what extent are multiword sequences associated with oral fluency? *Language Learning, 70*(2), 506-547. https://doi.org/10.1111/lang.12384

Therova, D. (2020). Review of academic word lists. *The Electronic Journal for English as a Second Language*, *24* (1), 1-15. https://www.tesl-ej.org/wordpress/issues/volume24/ej93/ej93a5/

Treffers-Daller, J. (2013). Measuring lexical diversity among L2 learners of French. In S. Jarvis & M. Daller (Eds.), *Vocabulary Knowledge: Human ratings and automated measures* (pp. 79-104). John Benjamins.

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

Wang, X. (2014). The relationship between lexical diversity and EFL writing proficiency. *University of Sydney Papers in TESOL, 9*, 65-88.

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., & Sasao, Y. (2013). New directions in vocabulary testing. *RELC Journal, 44*(3), 263-277. https://doi.org/10.1177/0033688213500582

Webb, S. (2021). The lemma dilemma. How should words be operationalized in research and pedagogy? *Studies in Second Language Acquisition, 43*(5), 941-949. https://doi.org/10.1017/S0272263121000784

Woolbert, C. H. (1922). Speaking and writing-A study of differences, *Quarterly Journal of Speech, 8*(3), *271-285.* *http://dx.doi.org/10.1080/00335632209379390*

Wu, S. Y., Huang, R. J., & Tsai, I. F. (2019). The applicability of D, MTLD, and MATTR in Mandarin-speaking children. *Journal of Communication Disorders, 77*, 71-79. https://doi.org/10.1016/j.jcomdis.2018.10.00

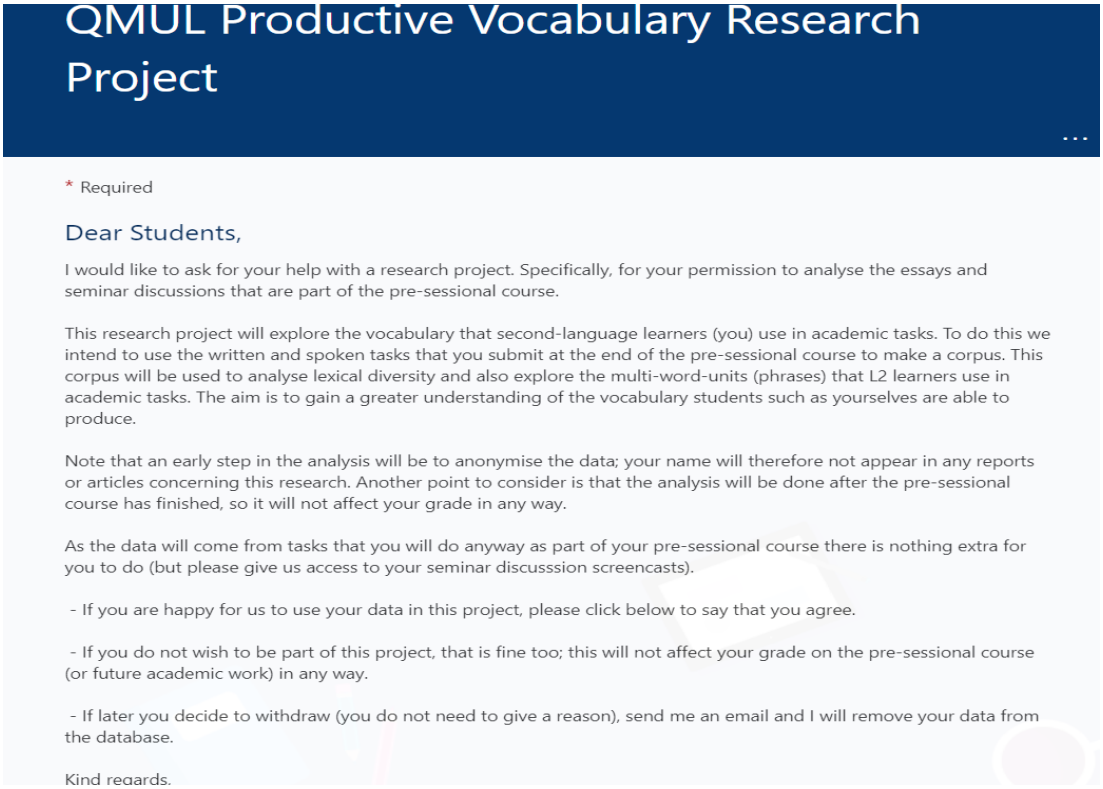
Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics, 31*(2), 236-259. https://doi.org/10.1093/applin/amp024

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing, 47*, 1-15. https://doi.org/10.1016/j.asw.2020.100505

Zhang, J., & Daller, M. (2019). Lexical richness of Chinese candidates in the graded oral English examinations. *Applied Linguistics Review*, *11*(3), 1-23. https://doi.org/10.1515/applirev-2018-0004

Appendices

Appendix 1: Written Consensus Form for the Written and Spoken Data Access (Chapters 3, 4, 5, and 6)



QMUL Productive Vocabulary Research Project

...

* Required

Dear Students,

I would like to ask for your help with a research project. Specifically, for your permission to analyse the essays and seminar discussions that are part of the pre-sessional course.

This research project will explore the vocabulary that second-language learners (you) use in academic tasks. To do this we intend to use the written and spoken tasks that you submit at the end of the pre-sessional course to make a corpus. This corpus will be used to analyse lexical diversity and also explore the multi-word-units (phrases) that L2 learners use in academic tasks. The aim is to gain a greater understanding of the vocabulary students such as yourselves are able to produce.

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.

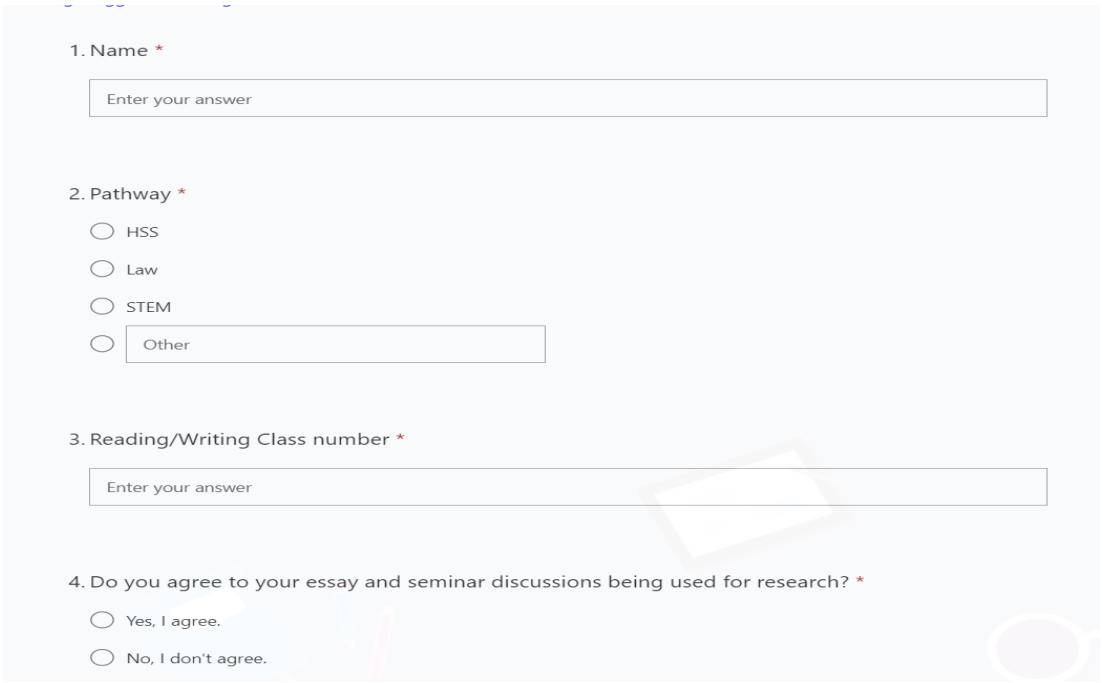
As the data will come from tasks that you will do anyway as part of your pre-sessional course there is nothing extra for you to do (but please give us access to your seminar discusssion screencasts).

- 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 pre-sessional 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.

Kind regards.



1. Name *

Enter your answer

2. Pathway *

- ◯ HSS
- ◯ Law
- ◯ STEM
- ◯ Other

3. Reading/Writing Class number *

Enter your answer

4. Do you agree to your essay and seminar discussions being used for research? *

- ◯ Yes, I agree.
- ◯ No, I don't agree.

Appendix 2: IELTS-Based Writing Band Descriptors (Chapter 3)

	80 – 100% CEFR: C2 / IELTS 8 - 9	70 – 79% CEFR: HIGH C1/IELTS 7-7.5	60 – 69% CEFR: LOW C1/IELTS 6.5 - 7	50 – 59% CEFR: HIGH B2/ IELTS 6.5 - 6	40 – 49% CEFR: LOW B2/ IELTS 5 – 6	30 – 39% CEFR: B1/IELTS 4 - 5	1 – 29% CEFR: A1/A2/IELTS below 4
Task Fulfilment [20]	• All parts of the question thoroughly addressed within a coherent argument. • Work shows main ideas prominently and clearly stated. • Introduction and Conclusion as in next band but also effectively show how work is related to question/topic. • Excellent presentation	• All parts of the question thoroughly addressed • Work shows good analysis. • Work is focused and only relevant issues presented. • Effective Introduction and Conclusion contextualise and draw ideas together, respectively. • Very good presentation throughout the paper	• All parts of the question addressed. • Work shows some ability to analyse • Work is focused and mainly relevant issues competently presented. • Good Introduction and Conclusion • Good presentation throughout the paper	• Addresses the question adequately. • Work shows an understanding of the topic but may be more descriptive than analytical. • There may be occasional loss of focus and irrelevancies in parts. • Introduction and Conclusion adequate. • Generally satisfactory presentation	• Parts of the question addressed but not all. • Work shows some understanding of the topic but is rather descriptive. • Some loss of focus & some irrelevancies may be evident. • Introduction and Conclusion may be rather simplistic with some inadequacies. • Presentation needs more care: some attempt to meet layout requirements.	• Not all parts of the question addressed. • Work shows weak understanding of the topic and is largely descriptive • Work is generally unfocused with many irrelevancies • Introduction and Conclusion may be simplistic and weak. • Presentation needs more care: some attempt to meet layout requirements but evident lack of proof-reading.	• Little attempt to address the question. • Work shows limited understanding of the topic • Work is unfocused and contains many irrelevancies • No, or extremely weak, Introduction and conclusion. • Poor presentation with little or no attention to layout
Organisation and Coherence [20]	Work shows completely logical organisational structure, enabling the writer's answer to the question to be followed effortlessly	• Good flow: ability to communicate with no difficulties for the reader. • Logical sequencing of ideas good text organisation, • Good paragraph organization, • Effective use of sign-posting expressions to create cohesion and coherence.	• Reasonably good flow: ability to communicate with few difficulties for the reader. • Good sequencing of ideas which enables the message to be followed clearly. • Paragraphs reasonably well organized although some room for improvement. • Fairly good use of sign-posting language to create cohesion & coherence	• Reasonable flow: ability to communicate, although with occasional difficulties for the reader. • Some ability to sequence ideas but overall structure may show some flaws. • Fairly good attempt to organize paragraphs into main and supporting ideas, some use of examples but insufficient. • An attempt to use sign-posting language, but sometimes inappropriate or inaccurate use; parts of the text may lack cohesion.	• Adequate flow: some ability to communicate but with some difficulties for the reader. • Some limits in the ability to sequence ideas and the overall organisation is likely to be flawed but the message may be followed adequately. • Paragraphs may lack unity but may show an attempt to use topic and supporting sentences. • Some attempt to use sign-posting language but it may be inappropriate in places; there may be some lack of cohesion.	• Ability to communicate but with strain for the reader. • The overall sequence of ideas may be flawed but the message can be followed in places. • Paragraph structure may be weak and disconnected – paragraphs may be only one or two sentences long and disjointed: little use of examples and illustrations. • Limited use of sign-posting language and often inappropriate; some lack of cohesion.	• Limited ability to communicate which often puts strain on the reader. • Ideas are poorly sequenced organized, and the message is difficult to follow. Lacks clear organization structure. • Paragraphs poorly organized and show little understanding of the purpose of paragraphs. • Very limited or inaccurate use of sign-posting language; lack of cohesion.

Referencing / Use of Sources [20]	• Excellent citation and full referencing in terms of accuracy. • Relevant and accurate incorporation of sources by summary, paraphrase, and quotation. • Sophisticated use of reporting language. • Bibliography is comprehensive and accurately reflects all sources cited	• Very good ability to select and reference a wide range of relevant sources correctly. • Good incorporation of sources by summary, paraphrase, and quotation. • Good use of reporting language. • Bibliography contains all sources referred to	• Good ability to select, incorporate and reference a reasonable range of sources adequately. • Shows some skill at incorporating sources by summary, paraphrase, and quotation. • Reasonable use of reporting language. • Bibliography contains most sources referred to	• Evidence of ability to select and reference sources but incorporation into text may be clumsy. • A reasonable attempt to incorporate sources by summary, paraphrase, and quotation • Fairly good attempt to use reporting verbs and expressions. • Bibliography may contain one or two omissions or inaccurate references.	• Some use of sources but the range may be limited and incorporation into the text is clumsy and may be without commentary. • Adequate attempt to incorporate sources by summary, paraphrase, and quotation. • Some attempt to use reporting verbs. • Some omissions evident in the Bibliography.	• Rather poor use of very limited sources and limited ability to incorporate them into the text. • Some attempt to incorporate sources by summary, paraphrase, and quotation but incorrect or lack of citation may result in plagiarism. • Limited use of reporting verbs • Bibliography shows several missing or incorrect references; may contain sources not referred to in the text.	• Inability to use source material. • Inability to summarise, paraphrase or quote which results in plagiarism. • Poor use of, or lack of, reporting language • Bibliography inadequate and inaccurate; there are likely to be a number of sources not referred to in the text.
Language Skills [40]	• Work demonstrates an authoritative use of the grammar and punctuation required for the task, ability to Manipulate complex structures • An excellent range of vocabulary appropriate to the task; completely accurate collocation and idiomatic expression • Excellent academic style, with appropriate use of register, ability to express caution and to generalize. • Absence of any errors indicates excellent proof-reading	• Work shows accurate grammar and punctuation, sophisticated sentence structures. • Good range of vocabulary appropriate to the task • Very good academic style with appropriate use of register, ability to express caution and to generalize. • Clear evidence of proof-reading.	• Work shows a good level of use of grammar and punctuation required for the task, some use of complex structures but perhaps incorrect use • Vocabulary generally appropriate to the task. • Good awareness of academic style (register, expression of caution, generalization). • Good evidence of proof reading but some errors may persist despite this.	• Work shows a reasonable use of grammar and punctuation with some ability to manipulate complex structures. There may be a limited number of grammatical errors, but these do not interfere with meaning. • Good range of appropriate vocabulary. • Awareness of academic style, but some inappropriacy in register, expression of caution may be weak and over generalizations may be evident. • Some lack of proof reading may result in careless mistakes,	• Work shows a basic grasp of grammar and punctuation but limited ability to manipulate complex structures. Errors may interfere with meaning. • Adequate range of appropriate vocabulary. • Some awareness of academic style but there are likely to be a number of over-generalisations and limited ability to express caution. • Inadequate proof reading may lead to careless mistakes.	• There may be recurrent grammatical and punctuation errors and limited ability to manipulate complex structures • Some inappropriate use of vocabulary. • Choice of style and register is often inappropriate. • Inadequate proof reading or lack of proof reading may result in careless errors.	• Significant, recurrent grammatical and punctuation errors. Very limited ability to manipulate structures appropriately and frequent errors in basic grammatical structures. • Range of vocabulary is inadequate for the task; errors make the meaning difficult to discern and cause strain for the reader. • Limited or no ability to use academic style • Lack of proof-reading results in incomprehension.

Appendix 3: Samples of L2 English Learners' 200-Word Essays of IELTS Writing 6.5, 7, and 7.5 Levels (Chapter 3)

Text 1: IELTS 6.5 Level Essay

"are the responsible of the rising connection and cited in and others and thanks to that convergence consumers share similarities between cultures and cited in and to this extend thanks to and open markets of neoliberalism the options to build your own personality trying to look like a citizen of other part of the world does not require much effort supporting the capability of choose there is the analysis of who assures that the plausibility of decide on yourself is the response to the massive shifting of fashion which provides no just goods but also with some particular ways to recreate you as well he supports that this has sense because of the rapid penetration of the capitalism in countries and the necessity of cover all the expectation of the consumers thus from a general perspective it has been argued that brings with it many movements that may involve political economic and even cultural in this sense it is also possible to observe that there is the possibility of choosing although it is true that possibility is remote in a certain way it is understood allowed to fight for a desire in other words possibility implies a kind of faculty"

Text 2: IELTS 7 Level Essay

"made a deal with countries that does not exist barriers when traveling, which means the citizens in could be quite easy to get along with abundant cultures in this way people could have a whole view of all kinds of cultures people must be more tolerant and open to accept different cultures and views validated that people even governments tend to enjoy peaceful lives which may decrease the possibility of armed conflicts in addition on the finally disintegrated the and the pattern collapsed at that time which meant that the global political form has become and eased to some extent the tense atmosphere between these countries about armed race has subsidized gradually it has helped the to become a super strong country in the world which may provide a relatively peaceful environment for the development of all both these examples has proofed that has reduced armed conflicts especially between the developed countries has broken up distance not just physically but also most dangerously mentally it created the illusion of intimacy when in fact the mental distances have changed little it has concerted the world without engendering the necessary respect recognition and tolerance that must accompany it itself is an exemplar"

Text 3: IELTS 7.5 Level Essay

“have appeared in the last years focusing the guiding principles in the context of addressing the relationship between business and human rights would be justified according to him this is because the effort by as the special representative of the including a long and comprehensive discussing process has obtained approval and therefore cements its position as the most universally dependable guideline in this field reaction to the guiding principles has been ranging from strong endorsement to severe criticism as of advocates observe that the guiding principles could help corporations to have higher levels of accountability and awareness in terms of the negative impact of business operations on human rights moreover they add that the concepts including due diligence process to prevent mitigate identify and account for their way to address their impacts on human rights contained in the guiding principles are appealing to companies because they make human rights manageable in fact the has welcomed the guiding principles and stepped up efforts to broaden their relevance as a key test for synergy between businesses and human rights and the quickly accepted and endorsed the guiding principles at its meeting moreover the energetically welcomed the approval of the guiding principles on”

Appendix 4: Samples of L1 Chinese L2 English Learners' 200-Word Essays of IELTS Writing 6.5, 7, and 7.5 Levels (Chapter 4)

Text 1: IELTS 6.5 Level Essay

"issue which is of the most important contents of global issues it can be said that with the existence and uniqueness of their homes and the simultaneous development of economic and between the various components of the different reactions in the earth life support system process are closely linked shall however due to irrational development because of the waste of natural resources global ecological damage pollution of all kinds more dangerous we all complain about the decline in the quality of life caused by environmental harm however each of us because of their little comfort every day to exacerbate this damage as far as financial development is concerned there are some differences in economic development between developed and developing countries economic is through trade direct foreign investment and multinational companies short term capital flows the international movement of workers and the general human and technology flows the national economy into the global economy from a historical perspective the significant growth in international trade over the centuries today exports and total imports of the countries up to of global production at the beginning of the nineteenth century this figure was less than and turning to developed countries financial development accelerate the"

Text 2: IELTS 7 Level Essay

"made a deal with countries that does not exist barriers when traveling, which means the citizens in could be quite easy to get along with abundant cultures in this way people could have a whole view of all kinds of cultures people must be more tolerant and open to accept different cultures and views validated that people even governments tend to enjoy peaceful lives which may decrease the possibility of armed conflicts in addition on the finally disintegrated the and the pattern collapsed at that time which meant that the global political form has become and eased to some extent the tense atmosphere between these countries about armed race has subsidized gradually it has helped the to become a super strong country in the world which may provide a relatively peaceful environment for the development of all both these examples has proofed that has reduced armed conflicts especially between the developed countries has broken up distance not just physically but also most dangerously mentally it created the illusion of

intimacy when in fact the mental distances have changed little it has concerted the world without engendering the necessary respect recognition and tolerance that must accompany it itself is an exemplar”

Text 3: IELTS 7.5 Level Essay

“it is serious that global companies overly employ the cheap labor of host as profit maximization as a result excessive exploitation of labor has impaired the rights of works in the host country and there are increasing problems for the vulnerable group for example the products of are sold in over countries which had become of the famous brands in the world however the sweatshop of in seriously damages the brand and reputation as they hired child labor and provided poor working conditions for workers after that still among the best due to that they take measures that donated dollars to the to implement a strict original equipment manufacturer policy and concern about the labor working conditions in factories which located in in terms of improving labor global companies have the responsibility to maintain the working environment and protect the basic right of workers in the host country the behavior that accord with business ethics is beneficial to improve the social economic conditions in developing countries and promote the development which also helps create stable international investment environment hence from the aspect of business ethics there is no doubt that has obvious effects on the behavior of corporations global corporations”

Appendix 5: Samples of L1 Chinese L2 English Learners' 200-Word Essays of IELTS Writing 6.5 Low and High Sub-Levels (Chapter 5)

Text 1: 6.5 Low-level Essay

"issue which is of the most important contents of global issues it can be said that with the existence and uniqueness of their homes and the simultaneous development of economic and between the various components of the different reactions in the earth life support system process are closely linked shall however due to irrational development because of the waste of natural resources global ecological damage pollution of all kinds more dangerous we all complain about the decline in the quality of life caused by environmental harm however each of us because of their little comfort every day to exacerbate this damage as far as financial development is concerned there are some differences in economic development between developed and developing countries economic is through trade direct foreign investment and multinational companies short term capital flows the international movement of workers and the general human and technology flows the national economy into the global economy from a historical perspective the significant growth in international trade over the centuries today exports and total imports of the countries up to of global production at the beginning of the nineteenth century this figure was less than and turning to developed countries financial development accelerate the"

Text 2: 6.5 High-Level Essay

"are shopping they would be more attracted by the brand which performs well in the process of producing in survey conducted by a small business consortium in of consumers say they prefer the company which engages actively and plays a positive role to improve community situation most of them say that they are less likely to buy products from unethical business even though the price is low enough the development of corporation cannot realize without the supporting by stakeholders customers and investors social responsibility matches the expectation from stakeholders customers investors and many other people who stand by the corporation business a survey conducted by the in proved that orientated corporations gain more market share than those did not consider social responsibilities according to an organization committed to informing companies about corporate social responsibility quotes a survey which concludes that the corporations which maintain a balancing

between stakeholders shows the growth rate and employment growth than the corporations which just emphasize on how to maximize profit and shareholders themselves so obviously it is necessary to understand and pay attention on social responsibilities in order to retain the good relationship with important stakeholders and catering the expectation of consumers and investors”

Appendix 6: IELTS-Based Speaking Band Descriptors (Chapter 6)

	80 – 100% CEFR: C2 / IELTS 8 - 9	70 – 79% CEFR: HIGH C1/IELTS 7-7.5	60 – 69% CEFR: LOW C1/IELTS 6.5 - 7	50 – 59% CEFR: HIGH B2/ IELTS 6.5 - 6	40 – 49% CEFR: LOW B2/ IELTS 5 – 6	30 – 39% CEFR: B1/IELTS 4 - 5	1 – 29% CEFR: A1/A2/IELTS below 4
Presentation Content [20]	• Purpose of presentation is clear, appropriate, and fully achieved. • Presentation is clearly focused, and only relevant issues presented. • Excellent research which is clearly demonstrated through illustrations and examples. • All source material cited. • Visual aids are designed to a professional standard in terms of layout, bibliography, and contents. • Very good analysis, synthesis, and application of research.	• Good flow; causes no difficulties for listener. • Logical sequencing of ideas good organisation of presentation. Very clear introduction and Conclusion. Questions invited • Good organization within sections. • Effective use of sign-posting expressions to create cohesion and coherence.	• Purpose of presentation is clear, appropriate, and largely achieved. • Presentation focused; issues presented are mainly relevant issues. • Appropriate research is demonstrated through illustrations and examples. • All source material cited, despite minor errors. • Generally clear and well-designed visual aids. Some evidence of proof reading but some errors may persist despite this. • Some evidence of ability to analyse, synthesise and apply research.	• Appropriate and adequately achieved purpose though may lack clarity. • May be occasional loss of focus and irrelevancies in parts. • Presentation shows some evidence of research and an understanding of the topic. • All source material is cited, though with some errors. • Generally satisfactory design of visual aids. Some lack of proof reading may result in careless mistakes. • Presentation may be more descriptive than analytical	• Purpose of presentation is appropriate but may not be entirely achieved. • Some loss of focus & some irrelevancies may be evident. • Presentation demonstrates evidence of adequate research and some understanding of the topic • Most source material is cited, though with frequent errors. • Adequately designed visual aids. Inadequate proof reading may lead to careless mistakes. • Presentation may be rather descriptive.	• Purpose of presentation may be unclear or inappropriate. • Presentation is generally unfocused with many irrelevancies. • Presentation demonstrates little evidence of research, weak understanding of the topic. • Some citation of source material. • Visual aids may provide inadequate support for the presentation. Inadequate proof reading or lack of proof reading may result in careless errors. • Presentation may be largely descriptive	• Purpose of presentation unclear or inappropriate. • Presentation is unfocused and contains many irrelevancies. • Presentation demonstrates no evidence of research and limited understanding of the topic. • Little or no citation of source material. • Visual aids non-existent or inadequate. Lack of proof-reading results in incomprehension. • Presentation may be entirely descriptive.
Presentation Structure [20]	• Excellent flow; causes no difficulties for listener. • Logical sequencing of ideas very good organisation of presentation. Excellent introduction and Conclusion. Questions invited. • Very good organization within sections. • Excellent use of sign-posting expressions to create cohesion and coherence.	• Good flow; causes no difficulties for listener. • Logical sequencing of ideas good organisation of presentation. Very clear introduction and Conclusion. Questions invited • Good organization within sections. • Effective use of sign-posting expressions to create cohesion and coherence.	• Reasonably good flow; causes few difficulties for listener. • Good sequencing of ideas which enables the message to be followed clearly. Good Introduction and Conclusion. • Reasonably good organization within sections although some room for improvement. Questions invited. • Fairly good use of sign-posting language to create cohesion & coherence.	• Reasonable flow causes occasional difficulties for listener. • Some ability to sequence ideas but overall structure may contain flaws. Reasonable Introduction and Conclusion. • Fairly good attempt to organize sections into main and supporting ideas, some use of examples but insufficient. Questions invited. • Attempt at sign-posting language; sometimes inappropriate or inaccurate; parts may lack cohesion	• Adequate flow but causes some difficulties for listener. • Limited ability to sequence ideas and overall organisation may be flawed but the message can be followed adequately. Introduction and Conclusion may be simplistic, overlong, or rushed. Questions not immediately invited. • Sections may lack unity but may show an attempt to use topic and supporting sentences. • Attempt at sign-posting language but it may be inappropriate; there may be some lack of cohesion.	• Lack of flow but causes strain for listener. • Flawed overall sequence of ideas but message can be followed in places. Introduction and Conclusion: simplistic, weak; do not correspond to body. Questions not invited. • Section structure may be weak and disconnected – sections short and disjointed, little use of examples and illustrations. • Limited use of sign-posting language and often inappropriate; some lack of cohesion.	• Absence of flow which often puts strain on listener. • Ideas are poorly sequenced organized, and the message is difficult to follow. Lacks clear organization structure. Little understanding of the purpose of introductions and Conclusions. Questions not invited. • Sections poorly organized and show little understanding of the purpose of structure. • Very limited or inaccurate use of sign-posting language; lack of cohesion.

Seminar Leadership [20]	• There is a totally clear task for seminar participants and the content is all highly focused and relevant. • The student clearly demonstrates a very high level of awareness of his/her audience. • The discussion is excellently controlled throughout. • The student gives a highly lucid summary of the discussion at its conclusion.	• There is a clear task for seminar participants and the content is focused and relevant. • The student demonstrates very good awareness of his/her audience. • The discussion is very well controlled. • The student gives a very good, lucid summary of the discussion at its conclusion.	• There is a fairly clear task for seminar participants and the content is mostly focused and relevant. • The student demonstrates good awareness of his/her audience. • The discussion is well controlled. • The student gives a good summary of the discussion at its conclusion.	• There is a task for seminar participants and the content is mostly relevant, but there may be some lack of clarity. • The student has satisfactory awareness of his/her audience. • An acceptable attempt is made to control the discussion. • The student gives a satisfactory summary of the discussion at its conclusion.	• There is a task for seminar participants, but it may not be presented clearly. Some of the content may lack focus and relevance. • The student may lack awareness of his/her audience. • The discussion may not be well controlled. • The student gives a summary of the discussion at its conclusion, but this may lack clarity	• There may be some confusion about the task for seminar participants. The content lacks focus and relevance. • The student lacks awareness of his/her audience. • The discussion is only just controlled. • The student gives a summary of the discussion at its conclusion, but this lacks clarity	• The task for seminar participants may be inappropriate, or unclear and is poorly explained. The content is unfocused and irrelevant. • The student has little or no awareness of his/her audience. • The discussion is not controlled. • The student fails to give a summary of the discussion at its conclusion or does this very poorly.
Language Fluency [20]	• Clear pronunciation all the time. • Very good, fluent command of language with almost no hesitations and excellent control of speed. • Excellent use of intonation and stress to convey stance and topic changes • Register always appropriate for this type of interaction. • Script independent; very confident and effective use of non-verbal communication (e.g., facial expressions, appropriate appearance).	• Clear pronunciation most of the time. • Good, fluent command of language with few hesitations and very good control of speed. • Very good use of intonation and stress to convey stance and topic changes • Register always appropriate for type of interaction. • Script independent; confident and effective use of non-verbal communication (e.g., facial expressions, appropriate appearance)	• Generally clear pronunciation. • Good, fluent production with some hesitations but good control of speed. • Generally good use of intonation and stress to convey stance and topic changes. • Register generally appropriate for type of interaction. • Generally, script independent; effective use of non-verbal communication (e.g., facial expression, appropriate appearance)	• Pronunciation is clear but there are some mispronunciations. • Speaks with a degree of fluency but with limited control of speed and some hesitations. • Reasonable use of intonation and stress to convey topic changes, but stance may not always be evident. • Register reasonably appropriate for type of interaction. • Often script independent; often effective use of non-verbal communication (e.g., facial expressions) and acceptably appropriate appearance)	• Pronunciation is generally clear enough to be understood despite a noticeable accent. • Can speak but with significant hesitation. May require a 'sympathetic' interlocutor. • Intonation and stress may only occasionally be used to convey stance or topic change. • Register is just appropriate; may sometimes be inappropriate for type of interaction. • Partly script independent; some limited awareness of non-verbal communication (e.g., facial expressions used effectively on occasion, fairly appropriate appearance)	• Mispronunciation sometimes makes communication difficult. • Hesitations can make communication difficult. Speed may be too fast or too slow. Often requires a 'sympathetic' interlocutor. • Stance and topic change not signalled with intonation and stress. • Register is often inappropriate for interaction. • Script dependent; little awareness of non-verbal communication (e.g., facial expressions sometimes inappropriate, fairly inappropriate appearance)	• Mispronunciation severely impedes communication. • Frequent hesitation or lack of control over speed severely impedes communication. Requires a 'sympathetic' and active interlocutor.' • Little control of intonation and stress. • Register is inappropriate for interactions. • Script dependent; poor awareness of non-verbal communication (e.g., inappropriate facial expressions and/or inappropriate appearance)
Language Accuracy [20]	• Student demonstrates mastery of the grammar required for the task; excellent ability to manipulate complex structures, • Excellent use of vocabulary which is appropriate to the task • Excellent academic style with totally appropriate use of register, very good ability to express caution and to avoid overgeneralizing. • Clear evidence of proof-reading (in visuals) and practice in presentation.	• Student demonstrates an authoritative use of the grammar required for the task; good ability to manipulate complex structures. • Good use of vocabulary which is appropriate to the task. • Very good academic style with appropriate use of register, good ability to express caution and to avoid overgeneralizing. • Clear evidence of proof-reading (in visuals) and practice in presentation.	• Student shows an above average level of use of grammar required for the task, some use of complex structures but perhaps incorrect use. • Good range of appropriate vocabulary. • Good awareness of academic style (register, expression of caution, few overgeneralizations). • Good evidence of proof-reading (in visuals) and practice in presentation but some errors may persist despite this.	• Student shows a reasonable use of grammar with some ability to manipulate complex structures. There may be a limited number of grammatical errors, but these do not interfere with meaning. • Vocabulary generally appropriate to the task. • Awareness of academic style, but some inappropriate register, expression of caution may be weak, and overgeneralizations may be evident. • Some lack of proof-reading (in visuals) and practice in presentation may result in careless mistakes.	• Student shows a basic grasp of grammar, but limited ability to manipulate complex structures. Errors may interfere with meaning. • Adequate range of appropriate vocabulary: a narrow range of simple language. • Some awareness of academic style but there are likely to be several overgeneralisations and limited ability to express caution. • Inadequate proof-reading (in visuals) and practice in presentation may lead to careless mistakes.	• There may be recurrent grammatical errors and limited ability to manipulate complex structures • Some inappropriate use of vocabulary. • Choice of style and register is often inappropriate. • Inadequate proof reading and practice may result in careless errors.	• Significant, recurrent grammatical errors. Very limited ability to manipulate structures appropriately and frequent errors in basic grammatical structures. • Range of vocabulary is inadequate for the task; errors make the meaning difficult to discern and cause strain for the reader. • Limited or no ability to use academic style • Lack of proof reading and practice results in incomprehension.

Appendix 7: Samples of L2 English Learners' 200-Word Spoken Transcripts of IELTS Writing 6.5, 7, and 7.5 Levels (Chapter 6)

Text 1: IELTS 6.5 Level Spoken Transcript

"countries like and made most of their low cost labor and opened their borders by attracting investors and at the same time they themselves had become international investors the integration into world markets a long term to experience a high speed economic development especially as we can see from the picture between and opens a global door so that the value of exports grew almost per year and in was eighth largest exporter in the world by flows into constitute over of the total in the world there is a source on bottom right as globalization has intensified many countries have gone from poor or struggling countries status to that of converging country which means that they have a similar development process to emerging countries they can see the picture and some countries around has a similar process the girls of is amazing but other developing countries had also experienced spectacular growth according to and challenge posted growth rate in of and proceptivity that not all the countries have benefited from globalization some least developed countries asked you in past situations there are some reasons why they cannot get economic development because globalization widens the gap in income according to"

Text 2: IELTS 7 Level Spoken Transcript

"structures of poor conscience it cannot pray or sustainable law in improving the present state of poor countries you want wider the gap between the rich and the poor rise the social inequality issue thorite the second is lack of the correct government integration in that way it may cause the human life exploitation and the mana play of resources by the people at the top of pyramid so I will use examples to explain these aspects the first is the inclusion program of for electronic technology companies this common news has indeed posted it the industry seeing demand developing countries especially in patch it only raised the living standards of middle class and they have done nothing to improve lives or for the poor people because these companies ignored the equal distribution of education resources and opportunities Indian and they also they have wrong they will hover around the cognize to the concept of people this for many companies they think people means people who are leaving a developing country and may become more detailed protentional market in addition to social structure and other influencing factors the market with corrector government integration may face the challenge of financial crisis"

Text 3: IELTS 7.5 Level Spoken Transcript

“a good education it is change and effective education it change not individual life it change the world so what we need to do to have an effective education I think what we have to do is to calculate how much money we need for effective education for individual and after that every countries need to have a commitment to put enough investment in education system I think the which even we have in a developed countries of in education system it is not good enough also and what happened in developing countries is absolutely unacceptable because we have many places where children are not going to school I thought let us move to our integration of kind of connection which we have in the process of globalization how it is possible to have an integration in the process of globalization if we do not have it human rights if we do not have a rule of law if we do not have a democracy and freedom if some people cannot understand what my freedom means its leads to problems for example what happened in when of the publisher and many say the protest because they cannot understand what is”