Doctoral Thesis

Exploring the Extent to Which Productive Vocabulary Knowledge Tasks Detect Changes

Hosam Gamal Saad Elmetaher

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

Hiroshima University

March 2022

Summary

This thesis investigates (i) the extent to which productive vocabulary knowledge tasks can detect changes over a short period of time for two different proficiency groups and (ii) how such a change might relate to aspects of speaking and writing abilities.

The thesis consists of four experiments (one cross-sectional and three longitudinal). All of the experiments have been administered by the researcher. The first experiment (cross-sectional, *n* = 105 A2 participants) compared a widely cited measure—the Productive Vocabulary Levels Test (PVLT)—with two tasks from Fitzpatrick and Clenton (2017): Lex30 and G_Lex. The first experiment reported that productive vocabulary knowledge is task-dependent for a group of A2 proficiency participants.

The second experiment (longitudinal, *n* = 100 A2 participants) measured participants' vocabulary knowledge with the same three productive vocabulary knowledge tasks from the first experiment at two time intervals: 0 and 3 months. The second experiment reported that productive vocabulary knowledge change varies according to task for a group of A2 proficiency participants.

The third experiment (longitudinal, *n* = 50 B2 participants) (i) measured participants' vocabulary knowledge with the same three productive vocabulary knowledge tasks from the first and second experiments but with an additional productive vocabulary writing task (the LFP), and (ii) explored the extent to which productive vocabulary knowledge change (as detected by different productive vocabulary tasks) might be consistent with infrequent words use in a writing task at 0 and 3 months. The third experiment reported that (i) productive vocabulary change varies according to task for a group of B2 proficiency participants; and (ii) infrequent words use in writing consistently relates to productive vocabulary knowledge task (the PVLT) score for the same group of B2 proficiency participants.

The fourth experiment (longitudinal, *n* = 45 A2 participants) explored potential relationships between the vocabulary knowledge elicited by the same three vocabulary tasks from the first, second, and third experiments, and aspects of speaking fluency at 0 and 3 months. The fourth experiment reported that aspects of L2 oral fluency inconsistently relate to productive vocabulary knowledge task scores for a group of A2 proficiency participants.

The four experiments reported in this thesis allowed me to make three main broad claims. First, the extent to which productive vocabulary tasks can detect productive vocabulary knowledge change might be based on implicational and developmental scales. The implicational scale-based change might relate to the quantitative change of specific aspects of productive vocabulary knowledge (e.g., form, semantic appropriateness, or grammatical accuracy) as detected by tasks with different elicitation characteristics. The developmental scale-based change might relate to the qualitative change of productive vocabulary knowledge (i.e., the ability to produce words in context and to perform contextually demanding tasks) as illustrated by Jiang's (2000) word development theory. This claim of productive vocabulary tasks' sensitivity to detect change based on two different perspectives might provide insights on how aspects of productive vocabulary knowledge construct change over time for groups of different proficiency level.

Second, the extent to which productive vocabulary tasks can detect consistent productive vocabulary knowledge change for two different proficiency groups (A2 and B2) might be task-dependent. G_Lex was the only task (when compared with Lex30 and the PVLT) that was able to detect consistent change between the two proficiency groups (A2 in Chapter 4 and B2 in Chapter 5) in two short-term (three-month) longitudinal studies. This claim of G_Lex sensitivity to detect consistent change might offer some useful insights to the discussion about how to assess various (e.g., pre-sessional) short-term (i.e., 12-week) language programs.

Third, the extent to which productive vocabulary tasks can detect productive vocabulary knowledge change on IELTS writing and speaking tasks might be task-dependent. The PVLT was the only task (when compared with Lex30 and G_Lex) that was able to detect the infrequent words change in the IELTS Writing Task 2 for a group of B2 participants. G_Lex, on the other hand, was the only task (when compared with Lex30 and the PVLT) that was able to detect infrequent words in Part 3 of the IELTS speaking test for a group of A2 participants. This claim of both the PVLT and G_Lex sensitivity to detect infrequent words in IELTS Writing and Speaking might offer some pedagogical implications to language teachers, skill-based (e.g., speaking and writing) courses, and test preparation (e.g., IELTS) programs.

In addition to the previous three main claims and their potential implications, the thesis developed different equivalent versions of two productive vocabulary tasks (a new version of Lex30, and two new versions of G_Lex). The newly developed versions are designed based on the original tasks' criteria, and might be of a great assist for future productive vocabulary knowledge studies.

Table of Contents

		Page
Chapter 1	**Introduction**	**17**
1.1	Receptive Vocabulary Knowledge-Productive Vocabulary Knowledge Relationship	17
1.2	Productive Vocabulary Knowledge Tasks	18
1.3	The Study	19
1.3.1	Investigating the Extent to which Different Productive Vocabulary Tasks Detect Change over Time (Chapters 2–7)	20
Chapter 2	**Literature Review**	**22**
2.1	Productive Vocabulary Tasks	22
2.1.1	Laufer and Nation (1995): Vocabulary Size and Use: Lexical Richness in L2 Written Production	23
2.1.2	Laufer and Nation (1999): A Vocabulary Size Test of Controlled Productive Ability	25
2.1.3	Meara and Fitzpatrick (2000): Lex30: An Improved Method of Assessing Productive Vocabulary in L2	29
2.1.4	Fitzpatrick and Clenton (2017): Making Sense of Learner Performance on Tests of Productive Vocabulary Knowledge	33
2.2	Vocabulary in Context	38
2.2.1	De Jong, Steinel, Florijn, Schoonen, and Hulstijn (2012): Facets of Speaking Proficiency	38

2.2.2	Clenton, De Jong, Clingwall, and Fraser (2021): Investigating the Extent to which Vocabulary Knowledge and Skills Can Predict Aspects of Fluency for a Small Group of Pre-intermediate Japanese L1 Users of English (L2)	41
2.2.3	Johnson, Acevedo, and Mercado (2016): Vocabulary Knowledge and Vocabulary Use in Second Language Writing	44
2.3	Productive Vocabulary Knowledge Development	45
2.3.1	Elgort (2018): Technology-mediated Second Language Vocabulary Development: A Review of Trends in Research Methodology	46
2.3.2	Housen et al. (2008): Investigating Lexical Proficiency Development over Time - The Case of Dutch Speaking Learners of French in Brussels	48
2.3.3	Crossley, Salsbury, and McNamara (2009): Measuring L2 Lexical Growth Using Hypernymic Relationships	52
2.4	The Current Study	56
Chapter 3	**Exploring Differences between Productive Vocabulary Knowledge Tasks**	**58**
3.1	Introduction	58
3.2	The Study	60
3.2.1	Measures	60
3.2.2	Participants	65
3.2.3	Methodology	66
3.3	Results	68

3.4	Discussion	69
3.5	Conclusion	73
Chapter 4	**Investigating the Extent to which Three Productive Vocabulary Knowledge Tasks Detect Change among an A2 English User Group**	**74**
4.1	Introduction	74
4.2	The Study	75
4.2.1	Measures	75
4.2.2	Participants	76
4.2.3	Methodology	77
4.3	Results	78
4.4	Discussion	79
4.5	Conclusion	79
Chapter 5	**Investigating the Extent to which Three Productive Vocabulary Knowledge Tasks Detect Change among a B2 English User Group**	**81**
5.1	Introduction	81
5.2	The Study	82
5.2.1	Measures	83
5.2.2	Participants	84
5.2.3	Methodology	85
5.3	Results	86
5.4	Discussion	87

5.5	Conclusion	89
Chapter 6	**Investigating the Extent to which Three Productive Vocabulary Knowledge Tasks Consistently Detect Aspects of Oral Fluency Change among an A2 English User Group**	**91**
6.1	Introduction	91
6.2	The Study	94
6.2.1	Measures	94
6.2.2	Participants	95
6.2.3	Methodology	96
6.3	Results	97
6.4	Discussion	108
6.5	Conclusion	112
Chapter 7	**Discussion**	**113**
7.1	Productive Vocabulary Tasks' Sensitivity to Detect Change	114
7.1.1	Productive Vocabulary Tasks' Sensitivity to Detect Change According to Task Characteristics	114
7.1.2	Productive Vocabulary Tasks' Sensitivity to Detect Change According to a Proposed Theory of L2 Lexical Development	120
7.1.3	Productive Vocabulary Tasks' Sensitivity to Detect Consistent Change	128
7.2	Productive Vocabulary Tasks' Sensitivity to Detect Change in Frequent Words Use in IELTS Writing and Speaking	130

7.2.1	Comparing the Vocabulary Produced in Response to the Productive Vocabulary Knowledge Tasks and the IELTS Writing Task 2 for B2 Participants	131
7.2.2	Comparing the Infrequent Words Produced in Response to the Productive Vocabulary Knowledge Tasks and the IELTS Speaking Tasks 2 for A2 Participants	137
7.2.3	Potential Implications for IELTS Writing and Speaking Courses	142
7.3	Conclusion	144
Chapter 8	**Conclusion**	**146**
8.1	Study Claims	150
8.1.1	Detecting Changes in Productive Vocabulary Knowledge	150
8.1.2	Detecting Consistent Changes in Productive Vocabulary Knowledge	151
8.1.3	Detecting Productive Vocabulary Knowledge Change in IELTS Writing and Speaking Tasks	152
8.2	Limitations and Future Research	152
Appendices		**154**
1	The Lexical Frequency Profile (LFP; Laufer & Nation, 1995; Chapter 2)	154
2	The Productive Vocabulary Level Test (PVLT; Laufer & Nation, 1999; Chapter 2)	155
3	Lex30 (Meara & Fitzpatrick, 2000; Chapter 2)	164
4	G_Lex (Fitzpatrick & Clenton, 2017; Chapter 2)	167

5	Samples of A2 Participants' Responses to Lex30 at 0 and 3 Months (Chapter 4)	170
6	Samples of A2 Participants' Responses to G_Lex at 0 and 3 Months (Chapter 4)	172
7	Samples of A2 Participants' Responses to the PVLT at 0 and 3 Months (Chapter 4)	174
8	Samples of B2 Participants' Responses to the Lex30 at 0 and 3 Months (Chapter 5)	180
9	Samples of B2 Participants' Responses to the G_Lex at 0 and 3 Months (Chapter 5)	182
10	Samples of B2 Participants' Responses to the PVLT at 0 and 3 Months (Chapter 5)	184
11	Samples of B2 Participants' Responses to the LFP at 0 and 3 Months (Chapter 5)	190

References **192**

Acknowledgments

I would like to thank my supervisor, Dr. Jon Clenton, for his invaluable direction and guidance throughout the completion of this thesis. Many thanks also go to Dr. Amanda Edmonds, and my colleagues at Hiroshima University for their kind help.

List of Tables and Figures

		Page
Table 2.1	Mean and Standard Deviations Percentage of Different Word Families for the Three Different Proficiency Groups	24
Table 2.2	Mean Scores, F-tests, and Total Score of the PVLT for the Four Different Proficiency Groups	26
Table 2.3	Correlations between 2,000, 3,000, UWL, and 5,000-word Levels of Four Different Versions of the PVLT	27
Table 2.4	Mean Frequency Profiles for Lex30	31
Figure 2.1	EVST (Yes/No) Test Scores Compared with Lex30 Scores	31
Table 2.5	Descriptive Statistics for the Seven Standard Measures (All Types, Content Types, U-All, U-Content, D-All, G-All, and G-Content)	49
Figure 2.2	Vocabulary Test Capture: Lex30, LFP. BFP, and G_Lex	36
Table 2.6	Descriptive Statistics for the Nine Class Measures (Nouns, G (N/N), G (N/All), Verbs, G (V/V), G(V/All), Adjectives, G (A/A), and G (A/All))	50
Table 2.7	Descriptive Statistics for the Six Frequency Measures (FF1, FF5, FR66, G-FF1, G-FR5, and G-FR66)	51
Table 2.8	WordNet Hypernymy Mean and Standard Deviations over the 52-week Study Period	54
Table 2.9	MTLD Values: Mean and Standard Deviations over the 52-week Study Period	55
Figure 3.1	PVLT (Laufer & Nation, 1999)	61
Figure 3.2	Lex30 (Meara & Fitzpatrick, 2000)	63

Figure 3.3	G_Lex (Fitzpatrick & Clenton, 2017)	64
Table 3.1	Comparison between PVLT, Lex30, and G_Lex	65
Table 3.2	Descriptive Statistics for the Three Productive Tasks	68
Table 3.3	Correlations between the Three Productive Tasks	69
Figure 3.4	Revised Fitzpatrick and Clenton's "Vocabulary Test Capture Model" to include Laufer and Nation's (1999) "Productive Vocabulary Levels Test"	72
Table 4.1	Changes between the Test Mean Percentage Scores at the Two Testing Times for A2 Participants	78
Table 5.1	Differences between the LFP and IELTS Writing Task 2	83
Table 5.2	Changes between the Test Mean Percentage Scores at the Two Testing Times for B2 Participants	86
Table 5.3	Significant Changes across Time as Measured by Lex30, G_Lex, PVLT, and LFP among Different Proficiency Levels	88
Table 6.1	Speaking in Relation to a Single Productive Vocabulary Task	93
Table 6.2	Fluency Measure Mean Scores for the Three Different Speaking Tasks	98
Table 6.3	Vocabulary Scores at Time 1 and Time 2	99
Table 6.4	Correlations between Vocabulary Test and Fluency Measures at Time 1_Part 1	100
Table 6.5	Correlations between Vocabulary Test and Fluency Measures at Time 2_Part 1	101
Table 6.6	Correlations between Vocabulary Test and Fluency Measures at Time 1 and Time 2_Part 1	102
Table 6.7	Correlations between Vocabulary Test and Fluency Measures at Time 1_Part 2	103

Table 6.8	Correlations between Vocabulary Test and Fluency Measures at Time 2_Part 2	104
Table 6.9	Correlations between Vocabulary Test and Fluency Measures at Time 1 and Time 2_Part 2	105
Table 6.10	Correlations between Vocabulary Test and Fluency Measures at Time 1_Part 3	106
Table 6.11	Correlations between Vocabulary Test and Fluency Measures at Time 2_Part 3	107
Table 6.12	Correlations between Vocabulary Test and Fluency Measures at Time 1 and Time 2_Part 3	108
Table 6.13	Correlations between Vocabulary Tasks and Fluency Measures at Time 1 and Time 2	110
Figure 7.1	Samples of Participants' Responses to Lex30, G_Lex, and the PVLT for A2 Level	116
Figure 7.2	Samples of Participants' Responses to Lex30, G_Lex, and the PVLT for B2 Level	116
Figure 7.3	Suggested Implicational Scale of the Three Productive Vocabulary Tasks	117
Figure 7.4	Chapters 4 and 5: Participants' CEFR Levels of Proficiency	118
Table 7.1	Change across Time as Measured by Lex30, G_Lex, and the PVLT for A2 Level Participants	118
Figure 7.5	Implicational Scale for A2 Participants	119
Table 7.2	Change across Time as Measured by Lex30, G_Lex, and the PVLT for B2 Level Participants	120
Figure 7.6	Implicational Scale for B2 participants	120

Figure 7.7	Jiang's (2000) "Three Stages of Word Development"	122
Figure 7.8	Productive Vocabulary Knowledge Tasks and Jiang's (2000; 2002; 2004) "Three Stages of Word Development"	123
Figure 7.9	Hypothesized Task Developmental Scale for A2 Participants in Relation to Jiang's (2000; 2002; 2004) "Three Stages of Word Development"	126
Figure 7.10	Hypothesized Task Developmental Scale for B2 participants in Relation to Jiang's (2000; 2002;2004) "Three Stages of Word Development"	127
Table 7.3	Detecting Productive Vocabulary Knowledge Change across Time as Measured by Lex30, G_Lex, and the PVLT for Chapters 4–6	129
Figure 7.11	Lexical Resource in IELTS Task Writing Band Descriptors (public version) (IELTS, 2021a)	132
Table 7.4	Correlations between the Three Productive Vocabulary Task Mean Scores and Number of Infrequent Words Produced in Response to the IELTS Writing Task 2 for Test 1 for a Group of B2 Participants	136
Table 7.5	Correlations between the Three Productive Vocabulary Task Mean Scores and Number of Infrequent Words Produced in Response to the IELTS Writing Task 2 for Test 2 for a Group of B2 Participants	136
Figure 7.12	Lexical Resource in the IELTS Speaking Band Descriptors (public version) (IELTS, 2021b)	138
Table 7.6	Correlations between the Three Productive Vocabulary Task Scores and Number of Infrequent Words Tapped by the IELTS Speaking Tasks in the Two Testing Times for A2 Participants	140

Table 7.7 Percentage of Infrequent Words (non-1k) in the Three Tasks of the IELTS Speaking Test for a Group of A2 Participants 141

Table 8.1 Experimental Chapters (3–6) 148

Chapter 1 Introduction

Lexical knowledge (both receptive and productive) is arguably the focus of all language teaching. Receptive knowledge, sometimes called "passive knowledge," represents the vocabulary we use in listening and reading. Productive knowledge, sometimes called "active knowledge," represents the vocabulary we use in speaking and writing.

1.1 Receptive Vocabulary Knowledge-Productive Vocabulary Knowledge Relationship

Both receptive and productive vocabulary knowledge might be highly related. As an example, at an early stage of development, what learners can recognize might also relate more closely to what they can produce (e.g., if learners can recognize the word "research" at an early stage of lexical development, then it is more likely that they can produce it). Melka (1997) considers productive vocabulary knowledge as being at one end of a continuum with receptive vocabulary knowledge at the opposite end, where knowledge of vocabulary items can be situated anywhere from "passive" (receptive) to "active" (productive). This receptive–productive continuum has been often cited in vocabulary research (e.g., Henriksen, 1999; Laufer & Goldestein, 2004; Palmberg, 1987; Schmitt, 2010) and remains strongly present in the research literature to this day.

However, this receptive vocabulary knowledge–productive vocabulary knowledge relationship might not be always straightforward (Elmetaher, 2021) mainly due to differences related to (i) size; (ii) assessment; and (iii) construct. First, there is an overall assumption that the size of receptive vocabulary knowledge (i.e., needed for listening or reading) might be bigger than productive vocabulary knowledge (i.e., needed for writing and speaking) (Laufer, 1998; Nation, 2001; Read, 2000; Webb, 2005; 2007). Not all that we can recognize will always comfortably relate to what we can produce, probably because we do not need to produce all the words that we can recognize.

Second, "productive vocabulary knowledge is believed to be more difficult to estimate and assess [than receptive vocabulary knowledge] and reliable tools to assess it are very scarce" (Miralpeix, 2019, p. 192). Nation and Webb (2011) identify at least two potential reasons for this difficulty. They note that tests of productive vocabulary knowledge (i) face more challenges in attributing credit for partial word knowledge than do tests of receptive vocabulary knowledge (p. 304) and (ii) may have difficulty in producing results that are representative of a learner's entire lexicon (e.g., tests that rely on prompts to elicit productive vocabulary run the risk of targeting lexical fields about which certain learners may not have prior knowledge) (pp. 200–201).

Third, unlike receptive vocabulary knowledge, most researchers consider productive vocabulary knowledge to be a multifaceted construct (Fitzpatrick, 2007; Fitzpatrick & Clenton, 2010; 2017). In recent discussions of the construct of productive vocabulary knowledge, Schmitt (2019) and Gu (2019) propose an extension of what is covered by productive vocabulary construct insofar as they suggest that productive knowledge concerns the encoding of communicative content. Schmitt suggests that productive vocabulary knowledge "involves knowing a lexical item well enough to produce it when it is needed to encode communicative content in speech or writing" (2019, p. 269). Gu (2019) suggests the need to add consideration of "appropriateness in use, because real competency in vocabulary…. entails the ability to know when to use what with whom in what context" (p. 273). Such multiple aspects of productive vocabulary construct might raise concerns about how productive vocabulary tasks might estimate and assess productive vocabulary knowledge.

1.2 Productive Vocabulary Knowledge Tasks

Productive vocabulary tasks vary and range from tasks that ask participants to provide single word associations (e.g., Lex30; Meara, & Fitzpatrick, 2000) to tasks that ask

participants to use word forms with semantic appropriateness and grammatical accuracy in context (e.g., Lexical Frequency Profile (LFP); Laufer & Nation, 1995). Such variety in productive vocabulary tasks with different characteristics and elicitation demands raises a need for further studies to investigate how each task (i) taps into the productive vocabulary knowledge construct, (ii) detects productive vocabulary knowledge change over time, and (iii) detects aspects of language production skills (speaking and writing) that change over time. This third point of individual productive vocabulary tasks to potentially detect changes in aspects of speaking and writing skills over time might offer some potential pedagogical implications to language teachers, skill-based (e.g., speaking and writing) courses, and test preparation (e.g., IELTS) programs.

1.3 The Study

This thesis offers a novel approach to investigating the extent to which different productive vocabulary tasks can detect changes over time. I focus on three productive vocabulary knowledge tasks: Lex30 is designed to elicit single words using word stimuli (Meara & Fitzpatrick, 2000); G_Lex is designed to elicit single words using sentence stimuli (Fitzpatrick & Clenton, 2017); and the Productive Vocabulary Levels Test (PVLT) is designed to elicit single predetermined words using sentence stimuli (Laufer & Nation, 1999).

The motivation for running three tasks in parallel comes from Chapelle (2006). Chapelle situates the need for inferencing of test constructs not only in relation to the broader "communicative language ability" framework (Bachman, 1990; Bachman & Palmer, 1996) but also for "defensible inferences to appropriate constructs" (p. 54). She suggests that testing should be conducted with more than a single method and, for reliable inferences to be made, performance should be both dependable and generalizable to the extent that "the inference […] assumes that performance on the test task generalizes to performance on other similar

tasks [and] also assumes that performance is dependable across tasks" (p. 55). Accordingly, Chapelle's paper highlights the need to concurrently employ multiple tasks that have been validated and shown to be reliable to infer the extent to which tasks elicit the overlapping or equivalent construct being investigated.

1.3.1 Investigating the Extent to Which Different Productive Vocabulary Tasks Detect Changes over Time (Chapters 2–7)

Chapters 2–7 have been developed to investigate the extent to which different productive vocabulary tasks detect changes over time. Chapter 2 presents the literature review and is divided into three sections. The first section presents a review of four studies, representing four different productive vocabulary tasks. The second section in this review discusses three papers' attempts to assess productive vocabulary knowledge in relation to aspects of speaking and writing skills. The third and final section collects strands of the discussions from sections one and two by discussing three papers that attempt to evaluate productive vocabulary knowledge development.

To provide a solid background to the following experimental chapters, Chapter 3 explores the extent to which three productive vocabulary tasks (Lex30, G_Lex, and the PVLT) capture contrasting and overlapping aspects of knowledge. Chapters 4 builds on Chapter 3 to investigate the extent to which the three productive vocabulary knowledge tasks detect change over time for a group of A2 proficiency level learners. Chapter 5 extends the discussion from Chapter 4 to investigate (i) the extent to which the three productive vocabulary knowledge tasks detect change over time for a higher proficiency (i.e., B2) group and (ii) whether the extent to which the three productive vocabulary knowledge tasks scores change might be consistent with aspects of writing. To obtain a picture of both productive language skills, Chapter 6 investigates the extent to which the three productive vocabulary knowledge tasks' scores might be consistent with aspects of speaking.

Following the experimental chapters, Chapter 7 reflects on (i) productive vocabulary knowledge tasks' sensitivity to detect change for the two different proficiency groups (A2 and B2); and (ii) productive vocabulary tasks' sensitivity to detect infrequent words change in aspects of IELTS speaking and writing tasks for the same two different proficiency groups.

Chapter 2 Literature Review

The literature review presented in this chapter is divided into three sections. The first section presents a review of four studies, representing four different productive vocabulary tasks. The second section in this chapter reviews three cross-sectional studies that attempt to assess productive vocabulary knowledge in relation to aspects of speaking and writing skills. The third and final section in this chapter reviews longitudinal studies that relate to productive vocabulary development. The literature review ends with a summary of the key points of all three sections and sets up the research questions for the investigations to follow.

2.1 Productive Vocabulary Tasks

The four reviewed papers in this first section detail the four productive vocabulary tasks used in the experimental chapters (3-7). Although all four tasks activate user knowledge/ability (Meara, 1997), each might possess different task characteristics and elicitation demands. Such differences in task characteristics and elicitation demands might reveal different information about the productive vocabulary knowledge construct.

The section starts with a review of Laufer and Nation's (1995) "Vocabulary Size and Use: Lexical Richness in L2 Written Production" in which they present a productive composition-based vocabulary task, the LFP. The second review is of Laufer and Nation's (1999) "A Vocabulary Size Test of Controlled Productive Ability" in which they present a controlled productive word completion task, the PVLT. The third review is of Meara and Fitzpatrick's (2000) "Lex30: An Improved Method of Assessing Productive Vocabulary in an L2" in which they present the word association task, Lex30. The final review is of Fitzpatrick and Clenton's (2017) "making Sense of Learner Performance on Tests of Productive Vocabulary Knowledge" in which they present a sentence completion task, G_Lex.

2.1.1 Laufer and Nation (1995): Vocabulary Size and Use: Lexical Richness in L2 Written Production

Laufer and Nation (1995) introduced a composition-based measure of lexical richness, the Lexical Frequency Profile (LFP) (Appendix 1). The LFP reflects "the ability to use a word at one's free will as free productive ability" (Laufer & Nation, 1999, p. 37).

The LFP was validated through an experiment of three groups of different proficiency participants (with group one being the lowest followed by group two, and then group three). Group 1 comprised 22 participants from a university in New Zealand and represented the lowest proficiency group. Group 2 (20 participants) and group 3 (23 participants) were from a university in Israel and represented the second and the third proficiency levels, respectively.

Participants were required to respond to two compositions with at least 300-word tokens in length for each. Compositions' topics varied but were mainly for general debatable topics (e.g., "It is always what you do not have as a child that is important to you as an adult. Agree or disagree with this statement" (Laufer & Nation, 1995, p. 320)). Compositions were corrected for misspellings and incorrectly used items, and then compared to vocabulary lists by computer software. The LFP score consisted of tallying the percentage of infrequent words (i.e., non-1k) a participant used in the two general composition questions.

The significant differences between the three groups were found to be proficiency-based, with the participants in group 1 (lowest proficiency group) using more-frequent words (i.e., 1k and 2k), Table 2.1 from Laufer and Nation (1995, p. 316). Laufer and Nation, therefore, claim that the LFP was able to show different changes between different proficiency groups.

Table 2.1

Mean and Standard Deviations Percentages of Different Word Families for the Three Different Proficiency Groups

	1st 1,000		2nd 1,000		UWL		Not in Lists	
	Comp1	Com2	Comp1	Com2	Comp1	Com2	Comp1	Com2
Group 1	86.5	87.5	71	70	32	41	33	28
SD	38	53	20	23	18	25	23	18
Group 2	79.7	79.4	67	68	81	78	56	66
SD	53	45	17	22	23	23	35	33
Group 3	77	74	66	56	81	101	75	87
SD	61	59	26	25	32	29	29	35
F-test	19.35	33.1	0.29	1.89	24.86	27.40	10,46	22.74
P-value	0001	0001	75	16	0001	0001	0001	0001

Note: Reprinted from "Vocabulary Size and Use: Lexical Richness in L2 Written Production," by Laufer and Nation, 1995, *Journal of Applied Linguistics*, 16(1), p. 316. Copyright 1995 by Oxford University Press.

There are at least three potential issues with the LFP. First, as Meara and Fitzpatrick (2000, p. 21) have pointed out, free productive vocabulary tasks such as the LFP might be "context-limited." Task-takers would have to write about specific given topics about which they may or may not be familiar. However, the effects of this limitation are usually minimized by using "a broad subject base" (p. 21), e.g., discussing common knowledge or general topics. Second, Meara and Fitzpatrick (2000) argued that it might be time-consuming to elicit two essays of 300 words each (as required by the LFP) from non-native speakers. I posit that this potential issue of being time-consuming might be even more challenging with

lower-level learners. Third, as Laufer and Nation (1999) argued, the LFP elicitation demands might not fully estimate learners' productive vocabulary knowledge. Learners might prefer to use more-frequent words over less-frequent ones when asked to respond to the LFP compositions. Thus, Laufer and Nation (1999) proposed "a constrained context" productive vocabulary level test that might be able to show learners' productive vocabulary knowledge at five different word-frequency levels as reviewed next.

2.1.2 Laufer and Nation (1999): A Vocabulary-size Test of Controlled Productive Ability

Laufer and Nation (1999) designed a controlled production measure of vocabulary, the Productive Vocabulary Level Test (PVLT) (Appendix 2). The PVLT was introduced for teachers to understand "the state of their learners' vocabulary knowledge" and for researchers to "draw on a variety of vocabulary measures to investigate the nature of vocabulary growth" (p. 33).

Laufer and Nation conducted two studies to investigate the PVLT's validity and reliability. In the first study, a total of 79 participants were divided into four different proficiency groups based on their class grade from lowest to highest: 10th grade, 24 participants; 11th grade, 23 participants; 12th grade,18 participants; and first-year undergraduate students,14 participants (Laufer & Nation, 1999, p. 38). In the second study, a total of 132 participants were divided into four different proficiency groups (n = 45, 36, 33, and 18). The second study participants were different from the first study.

In their first study, Laufer and Nation validated the ability of their PVLT task to distinguish among the four different proficiency groups. Accordingly, a full version of the PVLT was administered to the four groups. The PVLT consists of 90 selected words from five different frequency levels: 2k, 3k, 5k, UWL (University word list), and 10k. The first letters of each target word were provided in a sentence to restrict responses to a specific target word as an example of the task:

"The garden was full of fra_____ flowers."

The PVLT scores were processed as "correct/incorrect for each item" (p. 38). Simple spelling and grammatical mistakes were ignored. Six scores were given to each learner (five frequency levels plus a total score). Laufer and Nation found that the PVLT score changes as general proficiency increases (University > 12th grade > 11th grade> 10th grade), Table 2.2 from Laufer and Nation (1999, p. 39).

Table 2.2

Mean Scores, F-tests, and Total Score of the PVLT for the Four Different Proficiency Groups

	10th grade (*n* = 24)	11th grade (*n* = 23)	12th grade (*n* = 18)	University (*n* = 14)	F-test
2,000 level	11.8	15.0	16.2	17.0	17.9 p=.0001
3,000 level	6.3	9.3	10.8	14.9	21.2 p=.0001
UWL level	2.6	5.3	7.4	12.6	34.6 p=.0001
5,000 level	1.0	3.9	4.7	7.4	12.6 p=.0001
10,000 level	0.0	0.0	0.9	3.8	13.6 p=.0001
Total	21.7	33.4	40.1	55.8	32.6 p=.0001

* F-test results indicate that "the differences between the four groups of learners for the total scores and scores at individual frequency levels were significant." (Laufer and Nation, 1999, p. 40).

Note. Reprinted from "A Vocabulary-size Test of Controlled Productive Ability," by Laufer and Nation, 1999, *Journal of Language Testing*, 16(1), p. 39. Copyright 1999 by SAGE Journals.

In their second study, Laufer and Nation created three new versions of the PVLT, each with different items from the same five frequency levels (2k, 3k, 5k, UWL (University

word list), and 10k). The four versions of the PVLT were administered to four groups of learners. Each group responded to four different versions of a specific frequency level, e.g., a group took four different versions of the 2,000-word frequency level. An analysis of four of the five frequency levels (2k, 3k, 5k, and UWL) of the participants' responses found that the four versions of the task significantly correlated with each other (see Table 2.3 from Laufer & Nation, 1999, p. 43).

Table 2.3

Correlations between 2,000, 3,000, UWL, and 5,000-word Levels of Four Different Versions of the PVLT

	A/B	A/C	A/D	B/C	B/D	C/D
2,000 level (n = 45)	.82*	.82*	.78*	.83*	.81*	.77*
3,000 level (n = 36)	.71*	.70*	.82*	.82*	.71*	.80*
UWL level (n = 33)	.75*	.80*	.84*	.83*	.76*	.80*
5,000 level (n = 18)	.72 (p = .004)	.83*	.69 (p = .003)	.49 (p = .1)	.77 (p = .003)	.67 (p = .006)

*Significant at .0001 level.

Note: Reprinted from "A Vocabulary-size Test of Controlled Productive Ability," by Laufer and Nation, 1999, *Journal of Language Testing*, 16(1), p. 43. Copyright 1999 by SAGE Journals.

Laufer and Nation (1999) concluded that "the Productive Vocabulary Levels Test is a reliable, valid, and practical measure of vocabulary growth" (1999, p. 44). However, there are

at least three potential issues with the PVLT. First, the scoring responses for the three versions of the PVLT were taken from an outdated corpus (Nation, l983). An analysis of the 90 scoring responses of each version of the task with BNC-COCA (Nation, 2017) corpora revealed that the total number of scoring responses was different for the three different PVLT tasks: PVLT 1: BNC COCA total score = 85; PVLT 2: BNC COCA total score = 79; and PVLT 3: BNC COCA total score = 78.

Second, the PVLT scoring system might obfuscate the extent to which the task can detect knowledge. The task-taker is only given credit for responses that correspond to target items. As Walters (2012) has highlighted, a task-taker "might choose a different word to complete the [PVLT] sentence, with the further chance that this choice might be a less frequent word, possibly indicating a broader productive vocabulary than the test would reveal" (p. 173). In such cases, no credit would be given for knowledge of the less-frequent word.

Furthermore, Laufer and Nation assumed that the provided first letters of the scoring words should limit the possibility of using other words with semantical appropriateness but from different frequency bands. However, some of the PVLT sentences might arguably accept more than one possible response. As an example, from the University Word List level sentences in the first version of the PVLT:

"There has been a recent tr……. among prosperous families towards a smaller number of children."

The scoring response for this sentence is "trend"; however, other responses, such as "tragedy, track, trouble, trauma, transition" might arguably work.

Another example from the University Word List level sentences in the first version of the PVLT is:

"There are several misprints on each page of this te_____."

The scoring response for this sentence is “text”; however, other responses, such as “textbook, test, template” might arguably work.

Third, in relation to the earlier second point of the PVLT scoring system, Meara and Fitzpatrick (2000) argued that the scoring system of the PVLT of only one correct response for each task item might “identify what the learners do not know, but it is rather less successful at identifying the full extent of what they do know” (p. 21). They added that the PVLT might detect the knowledge of lower levels but not, for example, the 10,000-word level as it might be hard to assume that only 18 items can represent thousands of words.

In response to the PVLT scoring system’s ability to detect a change in higher levels, Meara and Fitzpatrick (2000) proposed a word-association productive vocabulary task, Lex30. They believed that Lex30 can produce adequate “rich vocabulary output” (p. 22) to better show learners’ productive vocabulary knowledge. Lex30 is reviewed in the next section.

2.1.3 Meara and Fitzpatrick (2000): Lex30: An Improved Method of Assessing Productive Vocabulary in L2

Meara and Fitzpatrick aimed to design an easy-to-use productive vocabulary task, Lex30 (Appendix 3). The authors argue that their newly developed productive vocabulary task is easy to administer, takes only a short time to complete, elicits adequate responses from the participants, and has the potential to be administered online.

To validate the Lex30, Meara and Fitzpatrick (2000) conducted a cross-sectional study with a group of 46 different L1 backgrounds. Their proficiency level was estimated by their class teacher as “high-elementary level to proficiency level” (p. 23). Participants were asked to complete the Lex30 along with a receptive proficiency “yes/no” English Vocabulary Size Test (EVST). Both vocabulary measures were completed within the same week.

Lex30 presents participants with a list of 30 English "stimulus words" taken from Nation's first 1,000-word list (Nation, 1984). Participants were requested to respond with up to four L2 words to each stimulus. Meara and Fitzpatrick (2000) selected cues based on three criteria: (i) high frequency from Nation's (1984) 1,000 frequency level words; (ii) do not elicit a primary response in comparison with first language (English) speaker data (Edinburgh Associative Thesaurus; Kiss et al., 1973); and (iii) do not elicit common responses in comparison with first language (English) speaker responses (Edinburgh Associative Thesaurus; Kiss et al., 1973).

In their study, Meara and Fitzpatrick provided the Lex30 cues one by one and asked participants to write at least three responses for each within 30 seconds. The 30 cues took about 30 minutes to complete. Full responses were lemmatized and then processed with software "similar to Nation's Vocal Profile" (Heatley & Nation, 1998). The software reports a full frequency-level profile for each participant. The Lex30 score consisted of a count of all but the highly frequent (i.e., non-1,000) responses.

Meara and Fitzpatrick found that most of the elicited words were within Nation's (1984) "first thousand" category. Table 2.4 is from Meara and Fitzpatrick (2000, p. 24). They also found, unlike second, third, and fourth responses, that the participants' first response was usually a frequent word.

Table 2.4

Mean Frequency Profiles for Lex30

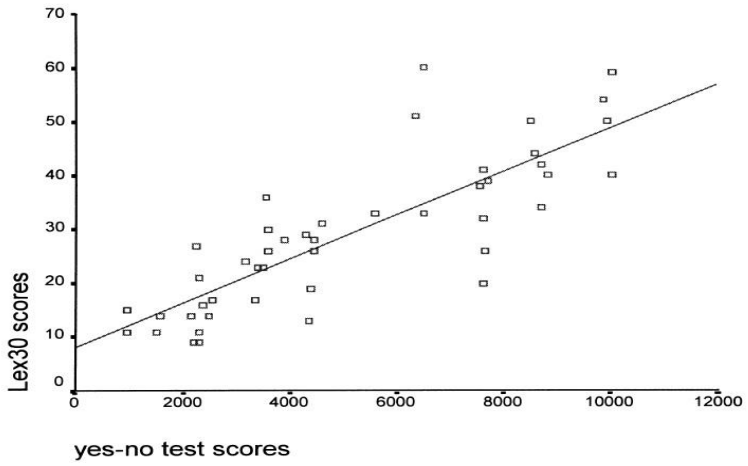
	Level 0	Level 1	Level 2	Level 3+	Total Words	Lex30
Mean	3.7	59.3	7.8	20.8	91.6	28.9
SD	3.6	13.9	3.6	11.4	24.2	13.9

Note: Reprinted from "Lex30: An Improved Method of Assessing Productive Vocabulary in an L2," by Meara and Fitzpatrick, T., 2000, *System Journal*, 28(1), p. 24. Copyright 2000 by Elsevier.

Meara and Fitzpatrick then compared Lex30 scores with the EVST vocabulary test scores as shown in Figure 2.1 from Meara and Fitzpatrick (2000, p. 25). They reported a significant correlation (0.841, $p < .01$) between both Lex30 and the EVST, indicating that both measures are quite predictive of one another.

Figure 2.1

EVST (yes/no) Test Scores Compared to Lex30 Scores



Note: Reprinted from "Lex30: An Improved Method of Assessing Productive Vocabulary in an L2," by Meara and Fitzpatrick, 2000, *System Journal*, 28(1), p. 25. Copyright 2000 by Elsevier.

Meara and Fitzpatrick interpreted these results as (i) Lex30 might be used as "a practical index of productive vocabulary" (p. 26) and (ii) Lex30 might be able to measure vocabulary knowledge differences.

Lex30, however, might possess a main potential issue. In an evaluation of Lex30, Walters (2012) introduced a sentence elicitation task to determine the extent to which "test takers could also use the words they were able to recall in association with the [Lex30] stimulus words" (p. 181). Based on her task elicitation responses, Walters suggested that Lex30 might elicit different types of productive vocabulary knowledge as a function of a learner's level of L2 competence, leading her to conclude that "Lex30 may be a valid test of productive vocabulary use for higher proficiency students, [whereas] it is more valid as a test of productive recall at the lower levels" (p. 183). Walter's point of Lex30 being "a valid test of productive vocabulary use" for advanced learners might be supported by Fitzpatrick and Meara's (2004) study in which they found a significant relationship ($p < 0.01$) between the PVLT and Lex30 for a group of intermediate to advanced proficiency level participants.

Walters' argument that Lex30 might be proficiency-based in eliciting productive vocabulary knowledge, on the other hand, might issue an urgent call of investigating productive vocabulary tasks elicitation abilities in relation to different proficiency levels. In a recent study that aimed to better understand productive vocabulary task production, Fitzpatrick and Clenton (2017) introduced (i) a vocabulary capture map to visually present task's ability to tap into the productive vocabulary knowledge construct and (ii) a new productive vocabulary task, G_Lex. The G_Lex was devised as a point of comparison to investigate the construct underlying Lex30. Their paper is reviewed in the next section.

2.1.4 Fitzpatrick and Clenton (2017): Making Sense of Learner Performance on Tests of Productive Vocabulary Knowledge

Fitzpatrick and Clenton (2017) attempted to investigate whether productive vocabulary knowledge tasks are assessing similar knowledge. The paper reported three separate studies in which the authors compared performance on four different productive tasks.

Fitzpatrick and Clenton included three groups of participants of 80, 80, and 100, respectively, all at a pre-intermediate to intermediate proficiency level. Participants were from different faculty majors.

The first study compared 80 participants' vocabulary scores on a word association task, Lex30 (Meara & Fitzpatrick, 2000; 2.1.3) to a composition task, the Lexical Frequency Profile (LFP; Laufer & Nation, 1995; 2.1.1). Both tasks were not significantly correlated (r = .186, p = .098). Despite adjusting LFP scores to reflect infrequent words in the same way as Lex30, the correlation between the LFP adjusted scores and the Lex30 scores remained insignificant (r = .108, p = .339). The authors highlighted the different task demands between the two tasks (Lex30 and LFP); e.g., LFP is a "discursive" task and might be eliciting more "function" words than Lex30.

In their second study, Fitzpatrick and Clenton devised a new Brainstorm Frequency Profile (BFP) task to remove composition writing demands. The BFP maintains the LFP question task, but elicits responses in single words as Lex30 does. The authors then compared a new sample of 80 students' BFP scores to the Lex30. The correlation between the scores (Lex30 and BFP) remained non-significant (r = .153, p = .175). Even after they modified the BFP point system to make it more similar to the Lex30, the tasks' correlation scores were still non-significant (r = .211, p = .061). The authors acknowledged that the non-significant correlation between both Lex30 and the BFP scores might indicate that the nonsignificant

correlation between Lex30 and the LFP in the first experiment might not relate to the LFP demands of producing vocabulary in context.

In their third study, Fitzpatrick and Clenton developed a new gap-fill vocabulary test (G_Lex; Fitzpatrick and Clenton, 2017). G_Lex (Appendix 4) is a sentence completion task in which up to five words are required to complete each of 24 sentence gaps (totaling a maximum of 120 possible responses, the same as for the Lex30 task). As with Lex30, any infrequent word provided by a respondent (i.e., any word that is not within the first 1,000 most frequent English words) receives one point. Moreover, words are scored as long as they "are spelled accurately enough to be identified" (Fitzpatrick & Clenton, 2017, p. 856). The 24 sentences were designed to elicit an identical number (8) of nouns, adjectives, and verbs, and each sentence met five criteria: (i) syntactically simple; (ii) consists of highly frequent words; (iii) potentially elicits five responses when trialed with first language (English) speakers; (iv) does not elicit sets of lexical (e.g., brown, blue, red); and (v) does not elicit similar responses in different sentences. To examine their assumption that the test performance difference highlighted in their first experiment might be one of "sampling" that might relate to "systematic differences in the quantity of elicitation prompts used" (p. 11), Fitzpatrick and Clenton (2017) compared 100 participants' scores of both Lex30 and their newly developed task (G_Lex). Both tasks' (Lex30 and G_Lex) correlation scores were significant ($r = .645$, $p = .01$).

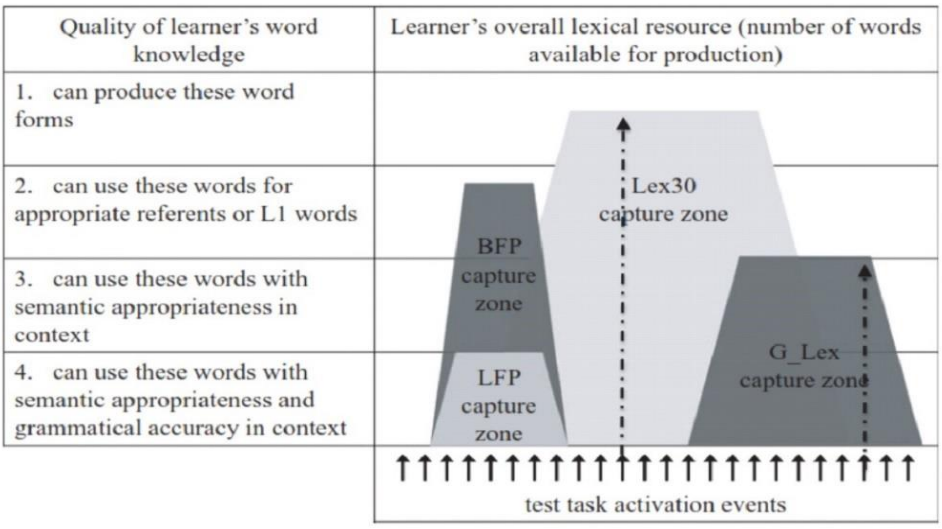
To make sense of such different results from their task comparisons and to compare "differences and similarities between test tools in a holistic and transparent way" (p. 862), Fitzpatrick and Clenton devised a "vocabulary test capture" model (2017, p. 860). They based their model on Paribakht and Wesche's (1993; 1996) Vocabulary Knowledge Scale (VKS). Originally, the VKS was devised to examine the vocabulary knowledge of 24 vocabulary items on a 5-point scale (i.e., from (I) "I don't remember having seen this word before" to (V)

"I can use this word in a sentence"). Rather than adopting a single scale, Fitzpatrick and Clenton's (2017) vocabulary test capture model adopts two scales or axes to "map" or interpret the productive vocabulary knowledge captured by different tasks. Their vertical axis maps the quality, or depth of word knowledge (see Figure 2.2) while the horizontal axis maps the quantity or breadth.

To demonstrate how users should interpret their model, Fitzpatrick and Clenton (2017) show, for example, that their newly devised G_Lex task likely captures the quality of knowledge at levels 3 and 4 (i.e., semantic as well as grammatical knowledge) in addition to multiple activation events (24 G_Lex sentences) suggestive of a relatively broad 'capture zone'. Their model serves to demonstrate that productive vocabulary tasks differ in terms of the extent to which the tasks require contextual knowledge, in addition to the number of conceptual activations. Fitzpatrick and Clenton's (2017) Test Capture Model shows that different tasks elicit productive vocabulary knowledge in different ways and that interpretation of the construct, therefore, appears manifestly multifaceted.

Figure 2.2

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



Note: Reprinted from "Making Sense of Learner Performance on Tests of Productive Vocabulary Knowledge," by Fitzpatrick and Clenton, 2017, *TESOL Quarterly Journal*, 51(4), p. 862. Copyright 2017 by TESOL International Association.

Fitzpatrick and Clenton's (2017) study, however, might possess two potential issues. First, their vocabulary capture map was based on task characteristics and not on how the task is being scored. For example, their newly developed G_Lex task was placed at levels 3 and 4 (i.e., semantic as well as grammatical knowledge). They assumed that the task score might reflect the ability to produce words with semantic appropriateness as well as potential grammatical accuracy. However, the authors indicated that responses are scored as long as they "are spelled accurately enough to be identified" (Fitzpatrick and Clenton, 2017, p. 856) regardless of their semantic appropriateness or grammatical accuracy. Thus, an interpretation of G_Lex scores might better place the task on level 2 (word meaning) of their vocabulary capture map instead of levels 3 (semantic appropriateness) and 4 (grammatical accuracy).

Second, more studies might be needed to validate Fitzpatrick and Clenton's newly developed G_Lex task. G_Lex was only validated by a single version of the task in a single study of 100 "pre-intermediate to intermediate" participants. Further studies with different proficiency levels and one or two more validated versions of the task might be needed.

To conclude, the earlier reviewed studies of the four productive vocabulary tasks (the LFP, the PVLT, Lex30, and G_Lex) reveal two important points. First, the four tasks make different claims that pertain to tapping into the construct of productive vocabulary knowledge. They might do so differently and, arguably, may provide insight into different aspects of productive vocabulary knowledge. The LFP is a composition-based vocabulary task. It might reflect the ability to produce words in context that are both semantically appropriate and grammatically accurate. The PVLT is a sentence completion task. It might reflect the ability to produce communicative content that is both semantically appropriate and grammatically accurate (Schmitt, 2019). Lex30 is a word association task. It might reflect the ability to produce partial knowledge of word forms, namely, knowledge restricted to form or to meaning. G_Lex is a sentence completion task. It might reflect the ability to produce semantically appropriate responses (Gu, 2019) that may or may not be grammatically accurate.

Second, the four tasks claim to discriminate between different proficiency levels. Laufer and Nation (1995) claimed that the LFP was able to discriminate change between different proficiency groups. Laufer and Nation (1999) claimed that the PVLT was able to discriminate change between three different proficiency groups. Meara and Fitzpatrick reported a strong correlation ($r = .0.84$, $p < .01$) between Lex30 and a receptive measure (EVST) score, leading them to argue that Lex30 might be "sensitive to gross differences in vocabulary knowledge" (p. 26). Fitzpatrick and Clenton (2017) reported a significant correlation between their newly developed productive vocabulary task (G_Lex) and Lex30

scores (r = .645, p = .01), leading them to argue that both tasks might be largely predictive of one another; thus, G_Lex might be as sensitive as Lex30 in measuring vocabulary knowledge differences.

Vocabulary tasks (e.g., PVLT and Lex30) have been used in earlier studies to estimate productive vocabulary knowledge in relation to aspects of language production skills, speaking, and writing. Some of those earlier studies (e.g., Clenton et al., 2021; De Jong et al., 2012; Johnson et al., 2016) are reviewed in the next section (2.2).

2.2 Vocabulary in Context

The three papers in this second section are reviewed to detail how the elicited knowledge by productive vocabulary tasks might relate to aspects of language production skills (speaking and writing). The section starts with a review of De Jong, Steinel, Florijn, Schoonen, and Hulstijn's (2012) "Facets of Speaking Proficiency" in which they explore possible relationships between vocabulary knowledge and second language speaking structure. The second review is of Clenton, De Jong, Clingwall, and Fraser's (2021), "Investigating the Extent to which Vocabulary Knowledge and Skills Can Predict Aspects of Fluency for a Small Group of Pre-intermediate Japanese L1 Users of English (L2)" in which they explore possible relationships between vocabulary knowledge and speaking fluency measures. The third and final review is of Johnson, Acevedo, and Mercado's (2016) "Vocabulary Knowledge and Vocabulary Use in Second Language Writing" in which they explore relationships between vocabulary knowledge and L2 writing performance.

2.2.1 De Jong, Steinel, Florijn, Schoonen, and Hulstijn (2012): Facets of Speaking Proficiency

De Jong et al. (2012) attempted to investigate different aspects of second language speaking proficiency. Their paper reported on a cross-sectional study in which they employed eight speaking tasks and six linguistic skills tasks.

De Jong et al. included 181 intermediate to advanced level second language learners and 54 native speakers of Dutch. The 181 L2 participants were aged 20–56 years (*M* age = 29, *SD* = 6). The 54 L1 speakers (control group) were aged 18–45 years (*M* age = 25, *SD* = 6).

The 14 speaking and linguistic skills tasks were presented to participants in two sessions. In the first session, eight speaking fluency computer tasks were administered. In the second session, two "linguistic knowledge" (vocabulary and grammar), three "linguistic processing skills" (reaction time measures and listening skills), and a "pronunciation skills" (speech sounds, word stress, and intonation) tasks were administered. The two sessions took about 2 hours for the L1 speakers and about 2.5–3 hours for the L2 learners to complete.

The speaking tasks aimed to measure different aspects of speaking fluency: (i) "complexity" (complex vs. simple topic); (ii) "formality" (informal vs. formal setting); and (iii) "discourse type" (descriptive vs. argumentative) (De Jong et al., 2012, p. 13). Responses to the speaking tasks were rated "in terms of the functional adequacy" by four judges on a six-level scale, totaling from 1 to 30 points.

The vocabulary task aimed to measure productive vocabulary knowledge and consists of two parts. The first part "elicited knowledge of single words" and was adapted from the PVLT (Laufer & Nation, 1999) in which 90 items (9 words X 10 frequency levels) were selected from the "Corpus Gesproken Nederlands" (CGN, Dutch Language Union, 2004). As in the PVLT, the first few letters of each item were presented in a "meaningful" sentence. The second part of the vocabulary task "elicited knowledge of multiword units" (p. 15). The second part followed the same structure as the first part but with 26 "prepositional phrases and verb–noun collocations" (p. 15). Both parts of the vocabulary task were processed with a maximum possible score of 118 points.

The grammar task aimed to measure different grammatical issues. Those grammatical issues relate to adjectives, verbs, main and sub-clauses, articles, pronouns, auxiliary verbs, and passive sentences. The grammar task was processed with a maximum possible score of 142 points.

The three linguistic skills tasks aimed to measure "lexical retrieval speed," "articulation speed," and "sentence building speed" (De Jong et al., 2012, p. 203). The listening task aimed to measure different aspects of listening skill (e.g., "vowels, diphthongs, consonants, intonation, and word-stress") (De Jong et al., 2012, p. 205). The four linguistic processing and pronunciation skills tasks were administered and processed through different software (e.g., "E-Prime software system") (Schneider, Eschman, & Zuccolotto, 2002a; 2002b), and PRAAT (Boersma & Weenink, 2005).

De Jong et al. used structural equation modeling (SEM) to explore different aspects of speaking proficiency. They reported two main findings. First, all linguistic skills, except two articulations "measures in the delayed picture naming task" showed significant relationships with "functional adequacy of speaking" (p. 28–29). Second, unlike Higgs and Clifford's (1982) assumption that language "components might change as L2 learners" proficiency increase over time. De Jong et al. found that different aspects of speaking proficiency were quite the same among two groups of different proficiency. De Jong et al. also reported that among functional adequacy language skills, vocabulary knowledge and pronunciation were the best "predictors" of speaking proficiency. They concluded that speaking fluency might be a matter of "declarative knowledge," "processing knowledge quickly," and "pronunciation skills." They suggested follow-up studies to investigate further factors that might affect speaking proficiency.

In partial replication of De Jong et al.'s (2012) study, Clenton et al. (2021) conducted a cross-sectional study in which they adopted a multifaceted approach to explore potential

relationships between a battery of tasks and different measures of fluency. Clenton et al. used the same three speaking tasks as those used in De Jong et al.'s (2012) study, but replaced their use of the PVLT with Lex30. Clenton et al. assumed that Lex30 might better relate to their pre-intermediate proficiency level participants (compared with the PVLT with intermediate to advanced level in De Jong et al.'s (2012) study).

2.2.2 Clenton, De Jong, Clingwall, and Fraser (2021): Investigating the Extent to which Vocabulary Knowledge and Skills Can Predict Aspects of Fluency for a Small Group of Pre-intermediate Japanese L1 Users of English (L2)

Clenton et al. (2021) attempted to (i) explore potential relationships between the knowledge elicited from a productive vocabulary knowledge task and the aspects of fluency elicited from three speaking (fluency) tasks adapted from De Jong et al. (2012); (ii) compare the vocabulary produced in response to the productive vocabulary task with the vocabulary produced in response to the fluency tasks; (iii) compare receptive knowledge with aspects of fluency; and, (iv) explore the speed and retrieval automaticity of the picture naming tasks in the investigation. Their paper reported on a cross-sectional small-scale study in which they employed various vocabulary knowledge and fluency elicitation tasks.

The study included 30 undergraduate adult L1 Japanese learners of English (*M* age = 19, *SD* = 1.3) with an average of 6.5 years of experience learning English. Participants X_Lex (receptive vocabulary task) scores (*M* = 4048, *Range* = 2400–4800) indicated a pre-intermediate proficiency.

Earlier studies on fluency (e.g., De Jong et al., 2012; 2.2.1) have shown strong and significant correlations between fluency measures and productive vocabulary knowledge. De Jong et al. (2012) explored fluency according to several fluency characteristics and reported strong and significant correlations between fluency and a newly constructed Dutch version of

a sentence completion task (Productive Vocabulary Levels Test (PVLT), Laufer & Nation, 1999).

Clenton et al. selected three speaking tasks as their productive vocabulary tasks from De Jong et al. (2012) (except Lex30; Meara & Fitzpatrick, 2000) along with an additional receptive vocabulary task (X_Lex; Meara & Milton, 2003). The three selected speaking tasks (from De Jong et al., 2012) possess different demands: (i) a formal descriptive task (e.g., describing a crime scene to a policeman; (ii) a formal persuasive task (e.g., responding in a town hall meeting to whether a new casino should be built next to an elementary school); and (iii) an informal persuasive task (e.g., responding to a view on climate change) (Clenton et al., 2021). To investigate different aspects of fluency (e.g., silent and filled pauses), the three speaking tasks were processed by PRAAT (Boersma & Weenink, 2005). Subsequently, the articulation rate was calculated per second of speaking time (total time minus total silent pausing time).

Lex30 (2.1.3) requires participants to respond with four words to each of the 30 cues. Following original task scoring procedures, Lex30 responses were processed online according to frequency using the Web VP. A Lex30 score consisted of a count of all but the highly frequent (i.e., non-1,000) responses.

X_Lex requires participants to respond to a “yes/no” task in which 120 words are presented and answers that indicate knowledge of the 1,000–5,000 frequency bands. X_Lex score is automatically processed with a maximum score of 5,000.

To further explore the extent to which vocabulary used in the speaking tasks can be predicted by measures of vocabulary knowledge and skills, Clenton et al. used three measures for vocabulary skills: LRS (Response Latency - picture naming), RL (Response Latency - delayed picture naming), and RD (Response Duration - delayed picture naming). Clenton et al. reported four main findings. First, in broad terms, the findings from their study might be

consistent with earlier fluency studies (e.g., De Jong et al., 2012) in that the productive vocabulary knowledge (as elicited by Lex30) might significantly correlate with the number of silent pauses. Second, unlike De Jong and Mora (2017), Clenton et al. did not find any significant relationships between receptive vocabulary knowledge task scores (as measured by X_Lex) and different aspects of speaking fluency. Third, two significant correlations related to speed and automaticity of retrieval were found: (i) a significant moderate correlation between response latency-delayed picture naming and the number of silent pauses per second in the speaking tasks ($r = .37$, $p < 0.05$); and (ii) a significant moderate correlation between response latencies in delayed picture naming and mean syllable duration ($r = -.44$, $p < 0.05$). Fourth, when correlating the vocabulary produced by both Lex30 and the speaking fluency task, an overlap was reported at levels 2 and 0 of the Academic Spoken Word List (ASWL; Dang et al., 2017).

To conclude, both Clenton et al. (2021) and De Jong et al. (2012) reported on a multifaceted approach in which they attempted to explore relationships between vocabulary knowledge and fluent speech. Both studies found significant relationships between productive vocabulary knowledge (as elicited by Lex30 in Clenton et al., 2021; and the PVLT in De Jong et al., 2012) and silent pauses. However, each study addressed specific proficiency levels (pre-intermediate in Clenton et al., 2021; and intermediate to advanced in De Jong et al., 2012). Thus, a possible partial replication might explore potential relationships between the knowledge elicited from multiple productive vocabulary tasks and tasks that elicit productive skills (e.g., speaking) for different proficiency level groups.

Regarding writing as the other productive language skill, the next reviewed study (Johnson et al., 2016) aimed to explore potential relationships between the vocabulary knowledge elicited by productive vocabulary tasks and L2 writing performance. As in De

Jong et al. (2012), Johnson et al. (2016) employed the PVLT with a group of advanced learners.

2.2.3 Johnson, Acevedo, and Mercado (2016): Vocabulary Knowledge and Vocabulary Use in Second Language Writing

Johnson et al. (2016) attempted to explore possible relationships between "vocabulary knowledge, vocabulary use, and L2 writing performance" (p. 702). Their paper reported on a cross-sectional study in which they employed different vocabulary knowledge tasks (Receptive vocabulary size test (Nation & Beglar, 2007); Aural vocabulary test (Fountain & Nation, 2000); Productive vocabulary levels test (Laufer & Nation, 1999); and a TOEFL independent writing task (ETS, 2005).

The study included 100 adult L2 English learners (*M* age = 19.58, *SD* = 3.53). Participants were enrolled in five different advanced-level classes at an educational institution in Lima, Peru.

The four tasks (three vocabulary tasks and one writing task) were administered by the classroom teacher with written instruction from the researchers with a one-week interval. Full responses were collected from 62 of the 100 participants. Both receptive and productive vocabulary tasks were processed and scored for each "correct" response. The aural vocabulary task was re-scripted and scored according to Fountain and Nation (2000). Scores were assigned for the "keywords." The TOEFL independent writing tasks were holistically assessed by two "experienced teachers" with a high interrater reliability rate between both (*r* = .98). The two essays were also processed with British National Corpus (BNC) for up to the 5k words.

Sets of correlations between the different tasks were then conducted. Johnson et al. reported that in general terms, the PVLT scores moderately correlated with the holistic

writing scores (r = .38, p = .003). This particular finding led Johnson et al. to argue that stronger L2 writers might know more of the infrequent words.

To conclude, as Clenton et al. (2021) and De Jong et al. (2012), Johnson et al. (2016) highlighted potential relationships between productive vocabulary knowledge tasks and contextual knowledge. The three studies reported on a multifaceted approach in which they explored relationships between vocabulary tasks and aspects of productive language skills (i.e., speaking fluency for Clenton et al., 2021 and De Jong et al., 2012; and L2 writing performance for Johnson et al., 2016). In addition, as in Clenton et al. (2021) and De Jong et al. (2012), Johnson et al. reported on a single proficiency level. Thus, a possible partial replication might explore potential relationships between the knowledge elicited from multiple productive vocabulary tasks and tasks that elicit productive skills (e.g., writing) for different proficiency level groups.

The two earlier sections (2.1 and 2.2) illustrated different productive vocabulary tasks and how they might relate to aspects of productive language skills (i.e., speaking and writing). However, the above seven reviewed studies were all cross-sectional and did not investigate how the knowledge elicited by productive vocabulary tasks might develop over time. The following studies (i.e., Elgort, 2018; Housen et al., 2008; Crossley et al., 2009) attempted to give an overview of how different aspects of productive vocabulary knowledge might develop over time for participants of different proficiency levels.

2.3 Productive Vocabulary Knowledge Development

This section starts with a review of Elgort's (2018) "Technology-mediated Second Language Vocabulary Development: A review of Trends in Research Methodology in Second Language Vocabulary Development (TMVD) Research" in which she presents a rich description of 82 selected studies from 2010 to April 2017 on technology-mediated vocabulary development. The second review is of Housen et al.'s (2008), "Investigating

Lexical Proficiency Development over Time – the Case of Dutch speaking Learners of French in Brussels" in which they attempt to investigate vocabulary knowledge development of a group of Dutch-speaking learners of French in a three-year study. The final review is of Crossley, Salsbury, and McNamara (2009), "Measuring L2 lexical growth using hypernymic relationships," in which they investigate vocabulary development in the spontaneous speaking of L2 learners.

2.3.1 Elgort (2018): Technology-mediated Second Language Vocabulary Development: A Review of Trends in Research Methodology

Elgort (2018) presented a meta-analysis of 82 selected studies from 2010 to 2017 on Technology-mediated Second Language Vocabulary Development (TMVD) research. Elgort reviewed the general methodological attributes of the 82 selected research papers. The majority of studies used university participants (n = 64), mainly university students. The total number of participants in each study varied, with an overall average of 67 and a standard deviation of 54. Out of the 82 studies, there were 11 studies with L1 Japanese participants and English was the main target language (n = 55).

Although teacher mediation was not included in most of the TMVD interventions, 60 studies were conducted in educational settings (e.g., class or a computer lab) and only 16 outside the formal educational settings by incorporating mobile-assisted language learning. The studies focused on the language receptive skills activities (reading: n = 16; listening: n = 11) were almost double the studies that focused on language productive skills activities (speaking and writing: n = 7 each). The target vocabulary words were between 5 and 156, with an average of 38 and a standard deviation of 34. However, there was no "item number" for some studies that did not focus on certain words or expressions. Interventions (e.g., language-focused activities, digital games, mobile-assisted language learning) ranged from a single class to months-long studies. Most studies (n = 50) used a pre- and post-test design

while 29 studies included a delayed post-test. Experimental or quasi-experimental designs were mostly used (n = 63) either exclusively or with other approaches. Quantitative data were mainly targeted with 50 studies and in combination with qualitative data in 28 studies. ANOVA, MANOVA, t-test, Mann-Whitney tests, or Wilcoxon tests were mainly used to analyze the quantitative data.

Upon analyzing the collected data, most of the reviewed studies (n = 67) investigated vocabulary development as the main focus; however, the measured aspects of word knowledge varied. The TMVD studies mainly measured participants' L2 lexicon quality (depth) of specific aspects of vocabulary knowledge rather than quantity (breadth) or size of the vocabulary knowledge. Out of the 82 reviewed studies, Elgort reported that vocabulary size, as a representative of vocabulary development, was investigated in only five studies, wherein only three studies investigated both vocabulary size and depth.

For further research designs, Elgort encouraged researchers to (i) "estimate and report participants' L2 proficiency and the measures used to estimate it" (p. 19), noting that a lack of sufficient information regarding general language proficiency (i.e., participants' L2 lexicon size) might cause difficulties for future replications, and (ii) devote more attention to studies that target fluency development. Elgort stated that out of the 82 reviewed studies, there was only one study "that explicitly emphasized fluency development" (p. 12). Elgort also noted that there is a shortage of TMVD speaking studies. Among the 81 selected studies, only one study listed "speaking" as a keyword. Thus, the following section aims to review two longitudinal studies (Housen et al., 2008 and Crossley et al., 2009) that attempt to investigate aspects of productive vocabulary knowledge development in L2 learners' speaking.

2.3.2 Housen et al. (2008): Investigating Lexical Proficiency Development over Time - The Case of Dutch Speaking Learners of French in Brussels

Housen et al. (2008) attempted to explore aspects of L2 lexical proficiency development over time. Their paper reported on a longitudinal study in which they investigated aspects of vocabulary development in French oral speech data: "unplanned oral retellings of the wordless picture story, 'Frog, Where Are You?'" (p. 7).

The study included 19 L2 French learners and 19 L1 French speakers from "ten different Dutch-medium secondary schools." The 19 L2 French learners had studied French for at least four years for an average of 150 minutes per week.

The "Frog" story task was administered three times over two school years (year 1 at age of 12 and year 3 at the age of 14) for the L2 French learners. The same task was administered only once (in year 2) for the L1 French speakers to provide a "benchmark" for the L2 French learners. In total, the "Frog" story task was administered four times, yielding a corpus of 76 speech datasets. Each speech dataset was recorded and transcribed, and then analyzed with CLAN software (MacWhinney, 2000).

Housen et al. calculated 42 different quantitative measures (only 22 were reported in this study) to determine the extent to which the quantity (i.e., number) and quality (i.e., type) of produced words by the L2 French learners would be similar to the quantity and quality of produced words by the L1 French speakers. The 22 reported measures were divided into three categories: standard (7 measures); class (9 measures); and frequency-based (6 measures). Housen et al. reported that all their seven "standard" measures showed significant development for their L2 French participants over the three collected data points (see Table 2.5 from Housen et al., 2008, p. 12).

Table 2.5

Descriptive Statistics for the Seven Standard Measures (All Types, Content Types, U-All, U-Content, D-All, G-All, and G-Content)

Measure	Year 1	Year 2	Year 3	Benchmark
#- All types	58.53	66.00	72.47	103.53
	(19.53)	(32.39)	(24.76)	(15.90)
#- Content types	28.89	35.89	39.79	56.32
	(9.98)	(18.62)	(14.25)	(9.45)
U-All	9.33	10.07	10.25	12.55
	(1.51)	(1.58)	(1.40)	(0.79)
U-Content	8.13	9.34	9.72	12.18
	(1.82)	(2.14)	(1.85)	(1.33)
D-All	15.80	18.39	19.56	33.39
	(7.90)	(8.63)	(8.18)	(5.79)
G-All	3.82	4.25	4.37	5.64
	(0.83)	(0.91)	(0.82)	(0.40)
G-Content	3.15	3.67	3.87	4.86
	(0.74)	(0.91)	(0.80)	(0.45)

*Numbers in parentheses are standard deviations.

Note: Reprinted from "Investigating Lexical Proficiency Development over Time - The Case of Dutch Speaking Learners of French in Brussels," by Housen et al., 2008, Journal of French Language Studies, 18(3), p. 12. Copyright 2008 by Cambridge University Press.

In addition, all nine "class" measures (except two adjective measures) showed significant development for the L2 French participants over the three collected data points (see Table 2.6 from Housen et al., 2008, p. 13).

Table 2.6

Descriptive Statistics for the Nine Class Measures (Nouns, G (N/N), G (N/All), Verbs, G (V/V), G (V/All), Adjectives, G (A/A), and G (A/All))

Measure	Year 1	Year 2	Year 3	Benchmark
Nouns				
# - Nouns	11.42	14.95	16.32	23.84
	(3.20)	(7.96)	(5.68)	(4.59)
G (N/N)	1.98	2.26	2.41	3.14
	(0.57)	(0.66)	(0.60)	(0.34)
G (N/All)	0.76	0.96	0.99	1.30
	(0.15)	(0.28)	(0.23)	(0.19)
Verbs				
#- verbs	15.37	18.16	20.37	29.42
	(6.29)	(9.48)	(7.85)	(5.50)
G (V/V)	2.36	2.69	2.82	3.55
	(0.57)	(0.68)	(0.62)	(0.48)
G (V/All)	1.00	1.17	1.22	1.60
	(0.30)	(0.33)	(0.28)	(0.23)

Adjectives				
#- Adjectives	2.11	2.79	3.11	3.05
	(1.52)	(2.25)	(1.82)	(1.58)
G (A/A)	0.95	1.20	1.27	1.18
	(0.45)	(0.66)	(0.43)	(0.46)
G (A/All)	0.13	0.17	0.19	0.17
	(0.08)	(0.12)	(0.08)	(0.08)

*Numbers in parentheses are standard deviations.

Note: Reprinted from "Investigating Lexical Proficiency Development over Time - The Case of Dutch Speaking Learners of French in Brussels," by Housen et al., 2008, Journal of French Language Studies, 18(3), p. 13. Copyright 2008 by Cambridge University Press.

Furthermore, the six "frequency" measures showed inconsistent development for their L2 French participants over the three collected data points (see Table 2.7 from Housen et al., 2008, p. 14).

Table 2.7

Descriptive Statistics for the Six Frequency Measures (FF1, FF5, FR66, G-FF1, G-FR5, and G-FR66)

Measure	Year 1	Year 2	Year 3	Benchmark
#-FF1	3.58	6.11	8.53	17.58
	(3.67)	(6.74)	(6.92)	(4.85)
#-FF5	10.63	13.05	14.42	20.89
	(6.76)	(14.47)	(11.66)	(7.09)
#-FR66	29.16	34.16	38.52	66.37
	(14.99)	(27.81)	(21.89)	(14.61)

G-FF1	0.22	0.35	0.49	0.96
	(0.23)	(0.30)	(0.32)	(0.25)
G-FR5	0.67	0.76	0.83	1.12
	(0.36)	(0.54)	(0.50)	(0.31)
G-FR66	3.91	4.41	4.80	6.40
	(1.15)	(1.50)	(1.28)	(0.60)

*Numbers in parentheses are standard deviations.

Note: Reprinted from "Investigating Lexical Proficiency Development over Time - The Case of Dutch Speaking Learners of French in Brussels," by Housen et al., 2008, *Journal of French Language Studies*, 18(3), p. 14. Copyright 2008 by Cambridge University Press.

Housen et al. concluded that although the quantity and quality of responses had developed within the three data collection intervals, development was higher between the first two times (Time 2 − Time 1) than the last two (Time 3 − Time 2).

To conclude, Housen et al. reported L2 lexical development in structured repeated oral tasks for intermediate to advanced level participants. To obtain a full picture of lexical development in speaking for different proficiency groups and, unlike Housen et al., Crossley et al. (2009) investigated L2 lexical development in spontaneous speech for lower-level learners.

2.3.3 Crossley, Salsbury, and McNamara (2009): Measuring L2 Lexical Growth Using Hypernymic Relationships

Crossley et al. (2009) attempted to investigate aspects of vocabulary development in spontaneous speaking of L2 learners. Their paper reported on a longitudinal study in which they examined the growth of hypernymic relations and of lexical diversity (a semantic

relationship between general (abstract, like “animal”) and specific (concrete, like “dog”) lexical items.

The study included six English learners enrolled in an academic English program at a university in the United States of America. The six participants were placed at the lowest proficiency level (level 1) out of a 6-level English program. The six participants were interviewed every two weeks over a full year.

The interviews were intended to elicit spontaneous speech in naturalistic settings. Participants interacted through different elicitation materials (e.g., questions, pictures, discussion topics). Participants were also encouraged to start their own topics. Each interview lasted 30–45 minutes. The interviews were recorded and then transcribed, yielding a total of 99 transcripts. Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004) software was used to measure both the quality (e.g., hypernymy values, word meaning values) and quantity (e.g., Textual and Lexical Diversity) aspects of learners’ lexical development over the study period.

Repeated measures ANOVA, linear curve, and two Pearson correlation tests were used to investigate relationships between study time and both hypernymic and lexical diversity development. Crossley et al. reported first that participants’ WordNet hypernymy values decreased over time: $F(5, 25) = 13.57$; $p < .001$; $\eta 2 = .73$ (see Table 2.8 from Crossley et al., 2009, p. 320).

Table 2.8

WordNet Hypernymy Mean and Standard Deviations over the 52-week Study Period

Week	Mean	SD (Standard Deviation)
2	1.44	0.22
4	1.34	0.18
16	1.23	0.15
32	1.19	0.13
50	1.02	0.08
52	1.05	0.11

Note: Reprinted from "Measuring L2 Lexical Growth Using Hypernymic Relationships," by Crossley et al., 2009, *Language Learning Journal*, 59(2), p. 320. Copyright 2008 by Wiley-Blackwell on behalf of the Language Learning Research Club at the University of Michigan.

Second, the authors found participants' MTLD (Measure of Textual Lexical Diversity) values increased over time: $F(5, 25) = 7.41; p < .001, \eta 2 = .60$ (see Table 2.9 from Crossley et al., 2009, p. 323).

Table 2.9

MTLD Values Mean and Standard Deviation over the 52-weeks Study Period

Week	Mean	SD (Standard Deviation)
2	28.43	7.27
4	25.37	4.55
16	32.26	7.27
32	31.12	3.78
50	34.88	4.25
52	35.43	2.92

Note: Reprinted from "Measuring L2 Lexical Growth Using Hypernymic Relationships," by Crossley et al., 2009, *Language Learning Journal*, 59(2), p. 323. Copyright 2008 by Wiley-Blackwell on behalf of the Language Learning Research Club at the University of Michigan.

Crossley et al. also found that different aspects of learners' lexicon developed over time. Learners tended to (i) produce more concrete words (i.e., hyponyms, such as "dog") at the initial stage of immersion and (ii) produce more abstract words (i.e., hypernyms, such as "animal") over time with an increase in L2 (English) study time. Such differences in the produced words' quality as the L2 study time increased might support the earlier claim in 2.1 that aspects of vocabulary knowledge might be proficiency-based. Crossley et al. also acknowledged some limitations for their study in which they suggested that further longitudinal studies should use a larger number of participants.

While the above two studies (Crossley et al., 2009; Housen et al., 2008) longitudinally described the development of quantitative and qualitative aspects of learners' lexicon, their findings were based exclusively on (i) a single vocabulary elicitation task ("interviews" for Crossley et al. and a structured "story task" for Housen et al.), which might not capture how

productive vocabulary knowledge develops in its various forms, and (ii) a small participant sample (n = 6 for Crossley et al. and n = 19 for Housen et al.).

2.4 The Current Study

Having reviewed some of the earlier vocabulary knowledge studies on productive vocabulary tasks, their use, and development, I now turn to outline the main points from all three earlier sections (2.1, 2.2, and 2.3) and close this section by setting up research questions for the investigations to follow.

Section 2.1 reviewed different productive vocabulary tasks (LFP, the PVLT, Lex30, and G_Lex) to see that each possesses different task characteristics. Although these tasks claim to (i) tap into the construct of productive vocabulary knowledge and (ii) be able to discriminate change, no study has been conducted to investigate the extent to which they might detect changes over time for groups of different proficiency levels.

Section 2.2 reviewed studies that highlight relationships between the knowledge elicited by vocabulary tasks and aspects of speaking and writing skills. Although the three studies reported significant relationships between vocabulary tasks and aspects of speaking and writing skills, they were mainly based on a single productive vocabulary task (for specific level of proficiency) and did not investigate if such significant relationships might remain consistent over time.

Section 2.3 reviewed productive lexical development studies. Crossley et al. (2009) and Housen et al. (2008) reported different aspects of lexical development in L2 speaking. However, their findings were based on a single productive vocabulary elicitation task for a small number of participants (n = 6 for Crossley et al. and n = 19 for Housen et al.). Concurrent studies of lexical development with multiple productive vocabulary tasks and a larger number of participants might be needed. Those concurrent studies might also explore different aspects of speaking (e.g., fluency) as suggested by Elgort (2018).

The design of the next experimental chapters is intended to build on and respond to gaps in these earlier reviewed studies in order to (i) investigate the ability of different productive vocabulary tasks to detect change over a short period for two different proficiency groups and (ii) provide implications for "global measures" (e.g., IELTS) in speaking and writing courses. More specifically, I employ three tasks of productive vocabulary knowledge (Lex30, G_Lex, and the PVLT) with a group of A2 learners and B2 learners in four studies (one cross-sectional and three longitudinal). Thus, in the next chapters, I attempt to answer the following research questions:

Q1. To what extent do results from a cross-sectional study suggest that productive vocabulary knowledge might be task-dependent?

Q2. To what extent do results from a longitudinal study suggest that three different productive vocabulary knowledge tasks might detect change for a group of A2 learners?

Q3. To what extent do results from a longitudinal study suggest that three different productive vocabulary knowledge tasks might detect change for a group of B2 learners?

Q4. To what extent do results from a longitudinal study suggest that three productive vocabulary knowledge tasks might detect changes in aspects of writing for a group of B2 learners?

Q5. To what extent do results from a longitudinal study suggest that three different productive vocabulary knowledge tasks might detect changes in aspects of speaking fluency for a group of A2 learners?

Chapter 3 Exploring Differences between Productive Vocabulary Knowledge Tasks

3.1 Introduction

Consensus suggests that eliciting productive vocabulary knowledge is far from a straightforward endeavor (e.g., Clenton, 2010; Fitzpatrick, 2007; Fitzpatrick & Clenton, 2010; 2017; Milton, 2009; Walters, 2012). Methods of assessment vary and tend to be based on the extent to which items have been carefully selected according to frequency (Ellis, 2002), or that they relate to specific contextual concerns or specific task demands. Tests based on frequency counts might be problematic for several reasons as frequency lists may not comfortably relate to items produced in writing or speaking (Milton, 2009). Context influences what learners produce in terms of sentence completion or composition tasks; because the tests might vary, other factors, such as test type and test function that elicit different task demands on test-takers, become the final issue to be addressed in the current study.

Three recent papers (Fitzpatrick, 2007; Fitzpatrick & Clenton, 2010; 2017) investigate whether productive vocabulary knowledge tasks are assessing similar knowledge. These studies suggest that research is equivocal in defining the construct of productive vocabulary. Comparisons between different productive vocabulary tasks reveal a mixture of findings. Fitzpatrick and Clenton (2010) report "that the tests are assessing broadly similar constructs" (p. 545) when comparing the Productive Vocabulary Levels Test (PVLT) (Laufer & Nation, 1999) and Lex30 (Meara & Fitzpatrick, 2000). Yet comparisons between two different productive vocabulary tasks, Lex30 and the Lexical Frequency Profile (LFP) (Laufer & Nation, 1995) report no statistical significance (Fitzpatrick & Clenton, 2017). Such findings corroborate Fitzpatrick and Clenton's (2010) earlier claim that "designing a test to address a

particular construct does not mean that it exclusively and comprehensively does so" (p. 545).

For the current discussion, I turn to Fitzpatrick and Clenton's (2017) three task comparisons devised to explore "differences and similarities between test tools in a holistic and transparent way" (p. 862). They begin by comparing Lex30 (a word association task) with the LFP (a composition task). Fitzpatrick and Clenton report that the correlation between the two tasks is not significant (r = .186, p = .098). Despite adjusting LFP scores to reflect infrequent words in the same way as Lex30, correlations remain non-significant (r = .108, p = .339).

Fitzpatrick and Clenton highlight the different task demands between the two (Lex30 and LFP) tasks and devise a new Brainstorm Frequency Profile (BFP) task to remove composition writing demands. The BFP maintains the LFP question task and elicits responses in single words as Lex30 does. Fitzpatrick and Clenton report non-significant correlations between the Lex30 and BFP tasks (r = .153, p = .175), and despite revising BFP (to reflect LFP) scoring, re-analysis scores remain non-significant (r = .211, p = .061).

For their third task comparison, they devise a G_Lex (Gapfill) task. Their G_Lex differs from their BFP task by presenting participants with "multiple activation events" rather than a single LFP question prompt as presented in their BFP task; the G_Lex task is like the LFP in the sense that the task requires test-takers to respond to context. A correlation analysis (r = .645, p <.01) between the Lex30 and G_Lex tasks suggests that the two are broadly predictive of each other.

To explore differences between these tasks, Fitzpatrick and Clenton devise a "vocabulary test capture" model (2017, p. 19). The two-dimensional model, based on Paribakht and Wesche's (1993; 1997) Vocabulary Knowledge Scale (VKS), maps the word knowledge elicited by different tasks of productive vocabulary knowledge. The vertical axis

maps the quality, or depth, of word knowledge that each task elicits. The horizontal axis maps the quantity, or breadth, of word knowledge that each task elicits (see Figure 2.2).

Fitzpatrick and Clenton (2017) suggest that the different axes of their model need investigating further regarding the conceptualization of productive vocabulary knowledge. The current chapter introduces an experiment to compare a widely cited vocabulary measure, the Productive Vocabulary Levels Test (PVLT) (Laufer & Nation, 1995; 1999) with others from Fitzpatrick and Clenton's (2017) model. Accordingly, the research question for the current chapter is:

RQ. To what extent do results from a cross-sectional study suggest that productive vocabulary knowledge might be task-dependent?

3.2 The study

To compare PVLT performance with the performance of other productive vocabulary tasks, the PVLT was administered with two tasks from Fitzpatrick and Clenton (2017): Lex30 and G_Lex. The three tasks are detailed below.

3.2.1 Measures

PVLT. The PVLT (Laufer & Nation, 1999; Figure 3.1) is a "controlled productive" word completion task and is cited widely (e.g., Fitzpatrick, 2007; Fitzpatrick & Clenton, 2010; Laufer & Nation, 1995; 1999; Meara & Alcoy, 2010; Read, 2012; Stæhr, 2009; Webb, 2009). On the PVLT, the first few letters of each target item are provided (between two and four letters) to restrict responses to a specific target word. The test elicits knowledge of five frequency levels (2k, 3k, 5k, UWL (the University Word List), and 10k) with 18 test sentences at each level considered to represent 1,000 words (the UWL represents 836 words). For Laufer and Nation (1999), the score one obtains on the PVLT corresponds to the total number of

correctly provided items; “mastery” of one level is a “matter of judgement [and] is probably around 15 or 16 out of 18 (85% or 90%)” (p. 41). This task was designed to allow task-takers to demonstrate the ability to encode communicative content and its authors, moreover, suggest that the PVLT may enable researchers to investigate the “developments [that] occur in the different types of vocabulary over a period of time” (p. 45).

A main criticism of the PVLT relates to the view that the task might obfuscate test-taker knowledge to the extent that a test-taker is only given credit for responses that correspond to target items. As Walters (2012) highlights, a “test-taker might choose a different word to complete the [PVLT] sentence, with the further chance that this choice might be a less frequent word, possibly indicating a broader productive vocabulary knowledge than the test would reveal” (p. 173). In such cases, no credit would be given for knowledge of the less-frequent word. Nonetheless, the PVLT remains a widely used measure for the assessment of learner productive vocabulary knowledge (e.g., De Jong et al., 2012).

Figure 3.1

The PVLT (Laufer & Nation, 1999)

1. I am glad we had this opp_____ to talk.
2. There are a doz_____eggs in the basket.
3. Every working person must pay income t_____.

Note: Adapted from “A Vocabulary-size Test of Controlled Productive Ability,” by Laufer, B., & Nation, P., 1999, *Journal of Language Testing*, 16(1), p. 46. Copyright 1999 by SAGE Journals.

Lex30. Lex30 (Meara & Fitzpatrick, 2000; Figure 3.2) is “basically a word association task,” reportedly “less constrained” than “context-limited productive tasks” (Meara & Fitzpatrick, 2000, p. 22). Lex30 has been widely used and appears in several recent research papers (e.g., Clenton, 2010; 2015; Fitzpatrick & Clenton, 2010; 2017; Fitzpatrick &

Meara, 2004; Fitzpatrick, 2007; Meara & Fitzpatrick, 2000; Walters, 2012). Lex30 requires participants to respond with up to four words to 30 stimulus words (totaling a maximum of 120 words). Carefully selected based on three criteria, each stimulus word: (i) is highly frequent and from Nation's (1984) 1,000 frequency level words; (ii) does not elicit a primary (predictable) response (e.g., dog, black) in comparisons with first language (English) speaker data (Edinburgh Associative Thesaurus; Kiss et al., 1973); and (iii) does not elicit common (frequent) responses in comparison with first language (English) speaker responses (Edinburgh Associative Thesaurus; Kiss et al., 1973). Completed Lex30 task papers are transcribed, lemmatized, and then compared with online corpora to decide a Lex30 score. All function words, proper nouns, numbers, and those words that fall within the first 1,000 frequency band do not score. A Lex30 score consists of a count of all but the highly frequent (i.e., non-1,000) responses. Lex30 scores are expressed as a simple count of the infrequent items. One strength of Lex30 is the ability for task-takers to showcase productive vocabulary knowledge activated by each of the 30 cues (Meara, 1997).

However, this ability to merely produce vocabulary items might hide important data. In an evaluation of Lex30, Walters (2012) introduced a sentence elicitation task to determine the extent to which "test-takers could also use the words they were able to recall in association with the [Lex30] stimulus word" (p. 181). Based on her elicitation task responses, Walters suggests that Lex30 may elicit different types of productive vocabulary knowledge as a function of a learner's level of L2 competence: "Lex30 may be a valid test of productive vocabulary use for higher proficiency students, [whereas] it is more valid as a test of productive recall at the lower levels" (p. 183). In discussing task differences with the PVLT, Walters suggests

that “Lex30 cannot be a replacement for a test like the PVLT” (p. 184) but reports “strong correlations” between the two.

Figure 3.2

Lex30 (Meara & Fitzpatrick, 2000)

1	attack				
2	board				
3	close				

Note: Adapted from “Lex30: An Improved Method of Assessing Productive Vocabulary in an L2,” by Meara, P., & Fitzpatrick, T., 2000, *System Journal*, 28(1), p. 28. Copyright 2000 by Elsevier.

G_Lex. G_Lex (Fitzpatrick and Clenton, 2017; Figure 3.3) was devised to investigate the construct underlying Lex30. G_Lex is a sentence completion task in which up to five words are required to complete each of 24 sentence gaps (totaling a maximum of 120 words, the same as for the Lex30 task). As with Lex30, any infrequent word provided by a respondent (i.e., any word that is not within the first 1,000 most frequent words in English) receives one point. Moreover, words are scored as long as they “are spelled accurately enough to be identified” (Fitzpatrick & Clenton, 2017, p. 856). The 24 sentences were designed to elicit the same number (8) of nouns, verbs, and adjectives, and each sentence met five criteria: (i) syntactically simple; (ii) consists of words from high-frequency bands; (iii) potentially elicits five responses when trialed with first language (English) speakers; (iv) does not elicit sets of lexical responses (e.g., brown, blue, red); and (v) does not elicit similar words in different sentences. G_Lex and Lex30 differ in their approach to eliciting word forms. Lex30 elicits single-word items from its single-word cues. G_Lex, however, uses sentence cues to elicit single-word responses. The two tasks are reported to be “moderately

predictive" of each other (2017, p. 856), with Fitzpatrick and Clenton reporting significant correlations ($r = .645$, $p < .01$).

Although G_Lex has not been as extensively tested as the two other measures used in the current study, G_Lex was retained for this study because its task properties situate it between the elicitation demands presented by both Lex30 and the PVLT. Unlike Lex30, G_Lex requires test-takers to consider the context (both in terms of part of speech and in terms of appropriate meaning) in providing their responses. However, G_Lex arguably is less "context-limited" than a task like PVLT in which one single response is expected. The three tasks (PVLT, Lex30, and G_Lex) have been summarized in Table 3.1.

Figure 3.3

G_Lex (Fitzpatrick & Clenton, 2017)

1.She loved to _____ over the phone.					
2.When I feel sad I always go to the_ _.					
3.They think car-racing is_____.					

Note: Adapted from "Making Sense of Learner Performance on Tests of Productive Vocabulary Knowledge," by Fitzpatrick & Clenton, 2017, *TESOL Quarterly Journal*, 51(4), p. 856. Copyright 2017 by TESOL International Association.

Table 3.1

Comparison between PVLT, Lex30, and G_Lex

Task	Elicitation Procedure	Number of Items	Scoring
PVLT	Fill in the blank - each item elicits one target word - the first few letters of the target word are presented	90 total - 18 items at 5 frequency bands (2k, 3k, 5k, University word list, 10k)	1 point for every target word correctly provided
Lex30	Word association	120 total - 30 items, up to 4 responses per item	1 point for every word form that is not within in the first 1k most frequent words
G_Lex	Fill in the blank (8 verbs, 8 nouns, 8 adjectives)	120 total - 24 items, up to 5 responses per item	1 point for every word form that is not within in the first 1k most frequent words

3.2.2 Participants

I assessed 105 (53 male, 52 female) second language (L2) undergraduate learners (of English) from a university in western Japan, all with L1 Japanese. All participants were aged 18 years and in their first year at university. Participants' TOEIC Bridge score ranged between 128 and 138, indicating a proficiency level of A2 according to the Common European Framework of Reference (CEFR) (ETS, 2019). All participants were placed in level 1 (lowest level of three) compulsory skills-based academic English classes (Listening I, Listening III, Reading I, and Active English I) for 90 minutes a week.

Participation in the study was voluntary with the right to opt out at any time. Personal data were protected as mainly the average mean test scores were reported in the study results, with no personal information collected from any of the participants.

3.2.3 Methodology

Testing was conducted over three weeks, with a week-long interval between each test. Classes comprised L1 Japanese students from different non-English majors and were, in effect, four different classroom groups. To negate the likelihood of any potential test effect, the three tasks (Lex30, G_Lex, and PVLT; Appendices 2-4, respectively) were presented in a different order to each classroom group.

Although all tasks awarded points only for infrequent (2k and less-frequent) words in accordance with their original scoring procedures, the sources used by the creators of each task to determine word form frequency were different. Responses reported in the original Lex30 paper (Meara & Fitzpatrick, 2000) were processed with Nation's (1984) word lists. Responses reported in the original G_Lex paper (Fitzpatrick & Clenton, 2017) were processed with JACET8000 word lists (Ishikawa et al., 2003). Correct responses to the original PVLT (Laufer & Nation, 1999) were taken from the (GSL+UWL) 2k–10k lists (Nation, l983).

To consistently evaluate responses to different productive vocabulary knowledge tasks, recent papers (e.g., Fitzpatrick & Clenton, 2017) used a single frequency benchmark to process participants' responses. Accordingly, Fitzpatrick and Clenton used Japan Association of College English Teachers (JACET) word lists that might better reflect their participants' (L1 Japanese) learning experience. Fitzpatrick and Clenton argued that using a single benchmark and identical scoring process for their participants' responses to the three productive vocabulary tasks (LFP, Lex30, and G_Lex) helped them to (i) compare learners'

performance across different tasks and (ii) map potential overlaps among the three tasks (refer to 2.1.4).

Following Fitzpatrick and Clenton (2017), the current study employs a single frequency benchmark (BNC-COCA; Nation, 2017) to compare responses to the different tasks. The combined British National Corpus (BNC) and the Corpus of Contemporary American (COCA) English represent the largest English language corpora. The BNC-COCA has been used in earlier studies with L1 Japanese participants (e.g., Laufer & McLean, 2016; McLean & Kramer, 2015). Dang (2020) suggested that BNC-COCA 2000 might better reflect learners' experience compared with other freely available corpora (e.g., New General Service List (New-GSL)). Accordingly, responses for the three productive vocabulary tasks (Lex30, G_Lex, and PVLT) were processed using the same online BNC-COCA (Nation, 2017) corpora. Task responses were first entered into individual computer text files. Scoring responses were from all frequency bands excluding those from the BNC-COCA 1k. I acknowledge that this approach represents a departure from the original PVLT scoring and stress that only the originally intended responses found to lie beyond the 1k frequency band in the BNC-COCA were accepted as scoring items for the PVLT.

In processing the PVLT with BNC-COCA, different versions of the PVLT generated different total numbers of scoring responses (i.e., PVLT (version A): BNC COCA total score = 85, not 90 as originally designed; refer to 2.1.2), whereas the two other tasks (Lex30 and G_Lex) elicited the same potential maximum of 120 items. To restate, both Lex30 and G_Lex raw scores consist of a count of all but the highly frequent (i.e., non-1,000) responses. Lex30 and G_Lex scores are expressed as a simple count of the infrequent items. To allow for comparison across the three tasks, each with a different number of maximum scoring items (i.e., PVLT = 85; Lex30 and G_Lex = 120 each), the participant mean raw scores were converted to percentages. The mean percentage scores represent (scoring items/possible

scoring items x100) on all three tasks. The scoring items refer to the number of infrequent (non-1k) responses for each task. The possible scoring items refer to the maximum number of infrequent words that each task can elicit. As an example, if a participant produces a total of 100 words in response to Lex30. Among those 100 words, only 25 are infrequent. This participant score on Lex30 is calculated as ((25/120) * 100) = 20.83%. The mean percentage scores of the three tasks (PVLT, Lex30, and G_Lex) were then statistically processed using a repeated measures analysis of variance (ANOVA).

3.3 Results

The results are presented in response to the following research question:

RQ. To what extent do results from a cross-sectional study suggest that productive vocabulary knowledge might be task-dependent?

Table 3.2 shows the descriptive statistics for the three different productive vocabulary task scores. Scores appear to vary according to the task: Lex30 score (15.341) is > PVLT (14.554) > G_Lex (7.651). Table 3.3 shows significant moderate correlations between the three tasks: Lex30 and PVLT scores ($r = .534$, $p < .001$); G_Lex and Lex30 scores ($r = .548$, $p < .001$); and G_Lex and PVLT scores ($r = .418$, $p < .001$), suggesting that tasks might be capturing overlapping aspects of knowledge.

Table 3.2

Descriptive Statistics for the Three Productive Tasks

	Mean	Std. Deviation	Range
PVLT	14.554	6.860	32.940
Lex30	15.341	8.884	34.160
G_Lex	7.651	3.874	15.840

Table 3.3

Correlations between the Three Productive Tasks

	PVLT	Lex30	G_Lex
PVLT	—		
Lex30	0.534***	—	
G_Lex	0.418***	0.548***	—

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

3.4 Discussion

To restate, Miralpeix (2019, p. 192) highlights that "productive vocabulary knowledge is believed to be more difficult to estimate [than receptive vocabulary knowledge], and reliable tools to assess it are very scarce." Thus, this chapter is focused on three different approaches (tasks) to assessing productive vocabulary knowledge. The ways these three tasks elicit productive vocabulary knowledge vary: Lex30 assesses the ability to produce individual items activated in response to single-word stimuli (Meara & Fitzpatrick, 2000); G_Lex assesses the ability to produce appropriate individual words activated in response to sentence word stimuli (Fitzpatrick & Clenton, 2017); and the Productive Vocabulary Levels Test (PVLT) assesses the ability to complete individual sentences when prompted with the first few letters of a specific predetermined item (Laufer & Nation, 1998).

Earlier studies (Fitzpatrick & Clenton, 2017; Fitzpatrick, 2007) attempted to consider task differences between the three tasks but independent of each other. The current study, therefore, builds on these earlier studies to investigate whether productive vocabulary knowledge varies according to the task. The experimental data appear to indicate significant differences in overall task mean scores for the three productive vocabulary task measures.

Differences in mean scores might relate to task characteristics. For instance, only minimal knowledge is elicited from the Lex30 task (or at least that the vocabulary knowledge

Lex30 elicits can only be demonstrated with knowledge of a word form). The G_Lex task elicits knowledge of context and, to some degree, semantics. There is more than one possible response to each of the G_Lex task's 24 cues and participant responses tend to fit the chosen grammatical context. The PVLT elicits knowledge of individual items that can only fit the specific grammatical gap available.

Correlations between the three tasks are moderately significant. Correlation between the PVLT and Lex30 scores ($r = .534$, $p < .001$) is similar to that reported in the earlier comparison by Fitzpatrick and Clenton (2010) ($r = .504$, $p < .01$) and suggests that the two tasks are "assessing broadly similar constructs" (p. 545). The correlation between G_Lex and Lex30 scores is also significant ($r = .548$ $p < .001$) but its effect size was not as strong as that reported in Fitzpatrick and Clenton (2017) ($r = .645$, $p < .01$) and might relate to differences in language proficiency. The correlation between G_Lex and PVLT scores was significant ($r = .418$, $p < .001$), suggesting that both tasks might be moderately predictive of one another.

The differences between the three tasks can perhaps more readily be explained with the modified vocabulary test capture model of Fitzpatrick and Clenton (2017). Therefore, I turn to Fitzpatrick and Clenton's (2017) paper whose aim was to better understand and interpret "Learner Performance on Tests of Productive Vocabulary Knowledge" by proposing a vocabulary test capture model. Their model was devised to compare "differences and similarities between test tools in a holistic and transparent way" (2017, p. 862) to advance understanding of what the tools measure. Fitzpatrick and Clenton included Lex30 and G_Lex in their original model, but not the PVLT. To support the discussion in the current chapter, I elaborate on Fitzpatrick and Clenton's model by adding the PVLT.

Fitzpatrick and Clenton (2017) describe their model as representing a "dual development trajectory" (p. 16), involving both "quality" (their vertical axis), which relates to whether individual tasks can capture different aspects of learner knowledge of individual

items, and "quantity" (their horizontal axis), reflecting the number of words known. More specifically, the y-axis is based on Paribakht and Wesche's (1993; 1996) Vocabulary Knowledge Scale (VKS) with a view to mapping the quality of depth of word knowledge elicited by different tasks of productive vocabulary knowledge. The y-axis is divided into four levels and should be understood as an implicational scale; for example, knowledge demonstrated at level 3 (semantic appropriateness) implies knowledge at both level 2 (knowledge of meaning) and level 1 (knowledge of form) for the word form in question. Fitzpatrick and Clenton assume that, with respect to knowledge elicited by different tasks, the higher the capture zone for a particular task, the broader the different types of partial word knowledge assessed in the task. The x-axis was intended to map the quantity, or breadth, of word knowledge elicited by each task by taking into consideration the number of activation events. In other words, tests that constrain learners to draw from a limited pool of items (by, for example, targeting a small number of items) will have a smaller footprint than tests that sample more broadly and, thus, presumably allow learners to showcase potential lexical knowledge from a larger number of semantic fields.

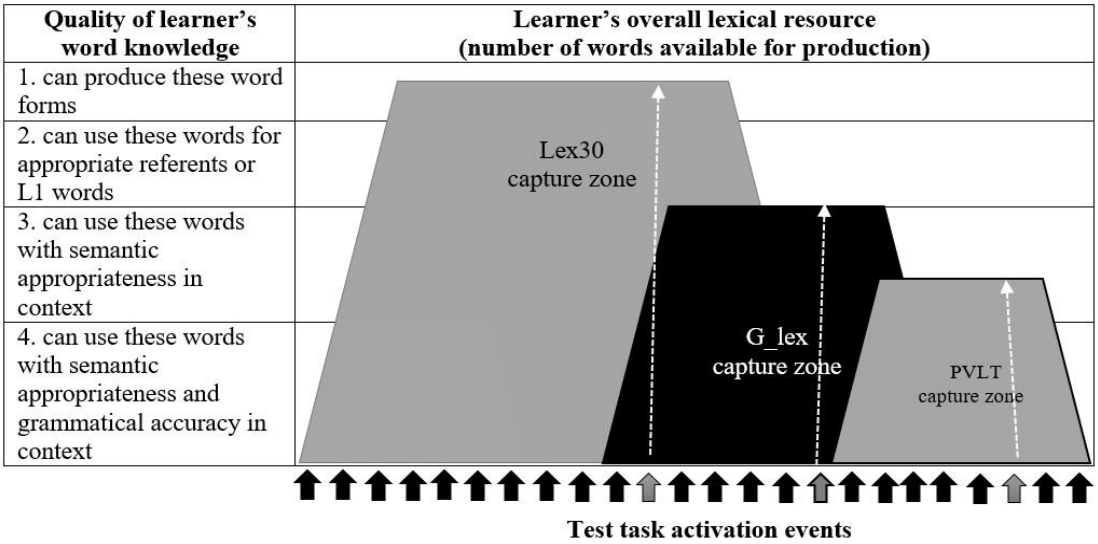
Utilizing this map, Fitzpatrick and Clenton (2017) compared several different tasks of productive vocabulary knowledge, showing, for example, that their newly devised G_Lex task likely captures the quality of knowledge at levels 3 and 4 (i.e., semantic as well as grammatical knowledge) in addition to multiple activation events (24 G_Lex sentences), which is suggestive of a relatively broad "capture zone" (see Figure 2.2). Lex30, on the other hand, is thought to crucially differ from G_Lex along the y-axis: given that Lex30 requires single-word responses in response to single-word cues, learners may demonstrate word knowledge at level 1 as well as levels 2-4. In addition, the two tasks also differ along the x-axis. Despite an equivalent number of activation events (i.e., 120 potential responses), I suggest that the G_Lex task likely elicits a smaller number of items than the Lex30 task

because G_Lex appears to elicit items that match the semantic appropriateness of the sentence elicitation cues, which I argue is likely more "context-limited" than having to provide word association responses as in the case of Lex30.

At this point, I turn to the question of how best to integrate the PVLT into this map. Unlike the Lex30 and G_Lex tasks, the PVLT only accepts one answer in response to each of its 90 sentence elicitation gaps, meaning that the range of activation events is more limited (and thus narrower along the x-axis). To provide a scoring response, learners must demonstrate knowledge of semantic appropriateness and grammatical accuracy (level 4) in context, meaning that the hypothesized capture zone for the PVLT is lower than that suggested for the Lex30 and G_Lex tasks (see Figure 3.4).

Figure 3.4

Revised Fitzpatrick and Clenton's "Vocabulary Test Capture Model" to Include Laufer and Nation's (1999) "Productive Vocabulary Levels Test"



Note: Adapted from "Making Sense of Learner Performance on Tests of Productive Vocabulary Knowledge," by Fitzpatrick and Clenton, 2017, *TESOL Quarterly Journal*, 51(4), p. 862. Copyright 2017 by TESOL International Association.

3.5 Conclusion

This chapter represents a first and new approach to evaluate vocabulary knowledge according to task. The chapter employed three different validated productive vocabulary knowledge tasks (Lex30, G_Lex, and PVLT) for a group of 105 L1 Japanese participants. The results indicate that (i) task scores appear to vary according to task (Lex30 > PVLT > G_Lex) and (ii) task performances appear broadly predictive of one another; for example, PVLT task scores broadly predict both Lex30 and G_Lex scores. To facilitate the discussion, the chapter proposes a visual representation of the results, thereby building on Fitzpatrick and Clenton's (2017) recent vocabulary capture map with the aim of exploring the extent to which the three different tasks capture contrasting and overlapping aspects of knowledge.

In the next chapter, I revisit Walters' (2012) suggestion by investigating the extent to which tasks might elicit different types of productive vocabulary knowledge as a function of L2 competence. The next studies, therefore, replicate the current study to explore the extent to which different productive vocabulary knowledge tasks detect change for different proficiency groups. Whereas a small body of previous research (e.g., Clenton, 2010; Fitzpatrick, 2007; Fitzpatrick & Clenton, 2010; 2017; Walters 2012) attempts to tease apart the relationship between tasks, constructs, and learner performance, I adopt a novel approach to investigate the extent to which different productive vocabulary tasks detect changes over time for two different proficiency groups. Thus, the following longitudinal studies allow me to hypothesize the relationship between tasks and productive vocabulary change.

Chapter 4 Investigating the Extent to which Three Productive Vocabulary Knowledge Tasks Detect Change among an A2 English User Group

4.1 Introduction

Chapter three outlined the hypothesized task characteristics of the three productive vocabulary tasks (Lex30, G_Lex, PVLT) under study using Fitzpatrick and Clenton's (2017) vocabulary task capture model. The results in Section 3.3 indicate that productive vocabulary knowledge might vary according to the task. The discussion in Section 3.4 details that the three productive vocabulary knowledge tasks represent three contrasting, but to some extent overlapping, scholarly interpretations of the productive vocabulary construct.

Building on Chapter 3 (Section 3.5), I longitudinally explore the extent to which different productive vocabulary knowledge tasks detect change for a group of A2 learners. The current chapter, therefore, replicates the experiment reported in Chapter 3 using the same three productive vocabulary tasks (Lex30, G_Lex, PVLT) with A2 learners level learners, but with an additional three-month longitudinal study with two testing times.

There are earlier examples of productive vocabulary test-retest (i.e., Fitzpatrick & Clenton, 2010). Fitzpatrick and Clenton (2010) argue that (i) "Lex30 produces consistent scores from learners over a short time period" (p. 549) and that (ii) Lex30 scores improved after a teaching intervention for six weeks for a group of 40 pre-intermediate L2 proficiency level learners. The current chapter is, therefore, re-employing Lex30 but with a longer period (12 weeks), a larger number of participants (n = 100), and most importantly, with two more productive vocabulary tasks (PVLT and G_Lex). Based on Fitzpatrick and Clenton's (2017) model, productive vocabulary knowledge is multifaceted; therefore, employing multiple tasks in a longitudinal study would help investigate how different productive vocabulary tasks detect change over time.

The three productive vocabulary tasks (Lex30, G_Lex, PVLT) (i) are predictable of each other (Chapter 3, Section 3.4) and (ii) appear to activate user knowledge/ability (Meara, 1997) although differences in the task characteristics (see Chapter 3, Section 3.2.1) are expected to detect different aspects of productive vocabulary knowledge. The Lex30 task has the potential to elicit all levels of partial word knowledge, namely, knowledge restricted to form or to meaning, and the G_Lex task was originally designed to examine the validity of Lex30. However, by examining participants' responses, it would appear that participants adhere to the parts of speech originally intended for the gaps; hence, the task appears to elicit an understanding of semantic appropriateness. The PVLT was "designed to elicit specific predetermined vocabulary items" (Fitzpatrick & Clenton, 2017, p. 846) and appears to elicit syntactic as well as semantic aspects of knowledge.

Accordingly, the research question for the current chapter is:

RQ. To what extent do results from a longitudinal study suggest that three different productive vocabulary knowledge tasks might detect change for a group of A2 learners?

4.2 The study

In this chapter, I report on a study that explores the ability of the three productive vocabulary knowledge tasks to detect change over time for a group of A2 learners.

4.2.1 Measures

The study employs the same productive vocabulary tasks from Chapter 3 (Lex30, G_Lex, and the PVLT, see Table 3.1). Because the experiment reported in this chapter is based on two test times, different equivalent versions of each task were required. On the basis that original versions were currently available for PVLT (3), Lex30 (2), and G_Lex (1), I developed one more Lex30 task and two more G_Lex tasks based on each original task creation criteria.

The new version of Lex30 (C, Appendix 3) was created using the same original task criteria as Meara and Fitzpatrick (2000). Accordingly, the Lex30 cues were selected as long as they (i) were highly frequent as per Nation's (2017) 1,000 frequency level; (ii) did not elicit a primary response in comparisons with first language (English) speaker data (Edinburgh Associative Thesaurus; Kiss et al., 1973); and (iii) did not elicit common responses in comparison with first language (English) speaker responses (Edinburgh Associative Thesaurus; Kiss et al., 1973).

The two G_Lex versions (B and C, Appendix 4) were created according to the original task criteria as Fitzpatrick and Clenton (2017). Accordingly, the 24 G_Lex sentences were designed to elicit an identical number (8) of nouns, adjectives, and verbs with each sentence: (i) being syntactically simple; (ii) consisting of only highly frequent words; (iii) eliciting five responses when trialed with first language (English) speakers; (iv) not eliciting lexical sets (e.g., brown, blue, red); and (v) not eliciting similar responses in different sentences.

4.2.2 Participants

I assessed 100 (60 males and 40 female) second language (L2) undergraduate learners (of English) from a university in western Japan, all with L1 Japanese. The participants in this study were different from the participants in the last study. All participants were aged 18 years and in their first year at university. Participants' TOEIC Bridge score ranged between 128 and 138, indicating a proficiency level of A2 according to the Common European Framework of Reference (CEFR) (ETS, 2019). All participants were placed in level 1 (lowest level of three) compulsory skills-based academic English classes (Listening II, Listening V, Reading II, and Active English II) for 90 minutes a week.

Along with their coursework materials, participants were provided with self-

study vocabulary activities. The self-study activities included a variety of academic vocabulary questions (e.g., true/false, multiple choice, sentence completion, matching, spelling). Participants were strongly encouraged to use the self-study vocabulary words in their different class activities.

Participation in the study was voluntary with the right to opt out at any time. Personal data were protected as mainly the average mean test scores were reported in the study results with no personal information collected from any of the participants.

4.2.3 Methodology

Participants took two equivalent versions of the three productive vocabulary tasks: Lex30 (Meara & Fitzpatrick, 2000), G_Lex (Fitzpatrick & Clenton, 2017), and the PVLT (Laufer & Nation, 1999) (Appendices 2-4). I conducted testing at two test points: at the beginning of their semester and at the end of the semester, three months later. To negate test effects, i.e., familiarity with the task's instruction and format, I presented the three tasks in different orders at each of the two test times.

As was the case in Chapter 3, individual responses to the productive vocabulary tasks were scored using the BNC-COCA online frequency corpora. A repeated measures ANOVA was conducted to explore the potential differences in the test mean percentage scores at the two time points. Samples of participants' responses are provided in Appendices 5–7.

As outlined in Chapter 3, percentage scores were used to address two main issues: (i) the different potential number of items for each task (because the PVLT elicits a potential maximum of 90 scoring items, whereas both Lex30 and G_Lex elicit a potential maximum of 120 scoring items); and (ii) the fact that different versions of the PVLT (Time 1 and Time 2), when processed with BNC-COCA, generate a different total number of scoring responses (the PVLT 1 (version B): BNC-COCA total score = 79); PVLT 2 (version C): BNC-COCA total score = 78), (Appendix 2). The mean percentage scores represent (scoring items/possible

scoring items x100) on all three tasks at the two time points. Table 4.1 presents the changes between the test mean percentage scores at the two testing times.

4.3 Results

Table 4.1 reports the different productive vocabulary task mean scores at the two different test times (Time 1, Time 2). Two-way repeated-measures ANOVA was used to examine the effect of Task and Time on the test mean scores percentage. There were significant main effects for both Tasks (F (2, 198) = 20.299, $p < .001$) and Time (F (1, 99) = 48.326, $p < .001$). Moreover, there was a statistically significant interaction between the effects of Task and Time (F (2, 198) = 11.797, $p < .001$). Post hoc comparison tests with Bonferroni correction were then performed to compare the mean differences across the two time points for each test. The post hoc tests showed that whereas scores on both Lex 30 and G_Lex significantly improved over time ($p < .001$), no such significant improvement was recorded for PVLT ($p = .718$).

Table 4.1

Changes between the Test Mean Percentage Scores at the Two Testing Times for A2 Participants

$N = 100$	Mean Score		
	Lex30	G_Lex	PVLT
Time 1	12.20 (7.33)	11.19 (5.67)	10.62 (5.99)
Time 2	16.35 (8.37)	13.30 (6.34)	11.32 (6.75)
Changes between time points			
Time 2 − Time 1	4.15***	2.11***	.7
	(5.69)	(5.42)	(4.77)

Note: * $p < .05$, ** $p < .01$, *** $p < .001$ (Bonferroni correction). Numbers in parentheses are standard deviations.

The findings from the study indicate significant changes over time for both Lex30 and G_Lex scores but not for PVLT scores among this group of A2 level participants.

4.4 Discussion

The introduction to this chapter outlined the need for a longitudinal investigation to explore the extent to which three different productive vocabulary knowledge tasks detect change for a group of A2 learners. The research question asked: To what extent do results from a longitudinal study suggest that three different productive vocabulary knowledge tasks might detect change for a group of A2 learners?

The research question facilitated an investigation that explored the extent to which productive vocabulary knowledge is being detected by the three different tasks that capture contrasting and overlapping aspects of knowledge. Although the three tasks claim to tap productive vocabulary knowledge, the use of longitudinal data potentially indicates that Lex30, G_Lex, and the PVLT are not consistently detecting the same kind of productive vocabulary knowledge. The findings indicate that Lex30 elicits a larger percentage of scoring items than the G_Lex and the PVLT tasks in both testing times (Time 1: Lex30 (12.20) > G_Lex (11.19) > PVLT (10.62) and Time 2: Lex30 (16.35) > G_Lex (13.30) > PVLT (11.32)). Both Lex30 and G_Lex scores show significant increases over time but not for PVLT scores among this group of A2 level participants (Time 2 − Time 1: Lex30 (4.15), G_Lex (2.11), and PVLT (.7).

4.5 Conclusion

This chapter investigated the extent to which three productive vocabulary knowledge tasks detect change with a group of A2 first language (L1) Japanese over a three-month longitudinal study. The results allow me to contribute to discussions detailing the kinds of information being detected by the three different productive vocabulary tasks that relate to productive vocabulary knowledge. Productive vocabulary knowledge change appears to

depend to some extent on the specific productive vocabulary task used. Both Lex30 and G_Lex scores appear to indicate some degree of significant change, but the PVLT scores do not.

Such inconsistent significant change among the three tasks might relate to proficiency. The participants' low English proficiency level (A2) might determine the extent to which they can respond to different contextual demands. Thus, a replication with different (higher) proficiency populations and an extra productive writing vocabulary task (LFP) might shed some light on the extent to which productive vocabulary tasks might (i) detect change for a higher proficiency group (higher than the A2 level reported in this study) and (ii) relate to aspects of writing.

Chapter 5 Investigating the Extent to which Three Productive Vocabulary Knowledge Tasks Detect Change among a B2 English User Group

5.1 Introduction

Chapter four outlined that productive vocabulary knowledge change appears to depend to some extent on the specific productive vocabulary task used for the group of A2 learners. The results reported in Section 4.3 indicate significant changes over time for both Lex30 and G_Lex scores but not for PVLT scores. The discussion in Section 4.4 indicates that the longitudinal data of the three administered productive vocabulary tasks might not consistently detect the same kind of productive vocabulary knowledge as Lex30 elicits a larger significant percentage of scoring items than G_Lex and PVLT at both testing times.

Section 4.5 suggests that such inconsistent significant change among the three tasks might relate to proficiency, which can be interpreted with respect to Meara's (1990; 1997) "network" metaphor. Milton (2013) explains: "[a] word that is recognized as a word in a language, but where nothing more is known about it, has no links and is not networked […] Words, which have this network of links around them can be said to be known more deeply than those, which are not known in this way" (p. 61). Thus, Walters (2012) suggests that lower-level learners may have less densely structured lexical networks. Higher-level learners, on the other hand, provide responses that they might be able to use in context and that may reflect their more densely structured networks. Following Milton (2013) and Walters (2012), I suggest that productive vocabulary knowledge task scores mean different things for different L2 proficiencies; hence, I might expect to see significant changes in written vocabulary knowledge, as measured by the lexical frequency profile (LFP) task, with learners from a different proficiency level than those reported in Chapter 4.

Unlike Lex30, G_Lex, and the PVLT, the LFP is based on composition. The LFP has been used in earlier cross-sectional and longitudinal studies with beginner level learners (e.g.,

Fitzpatrick & Clenton, 2017; Horst & Collins, 2006). Fitzpatrick and Clenton (2017) report a non-significant correlation for the LFP and Lex30 in a cross-sectional study with pre-intermediate L1 Japanese learners. Horst and Collins (2006) report inconsistent decline in their participants' infrequent written productive vocabulary knowledge when using the LFP (in response to picture prompts) in a longitudinal study with beginners L1 French.

To obtain a more complete picture of the development of infrequent written vocabulary knowledge, both studies (Fitzpatrick & Clenton, 2017; Horst & Collins, 2006) suggest follow-up studies with learners from different levels of L2 proficiency. Therefore, Chapter 5 replicates Chapter 4 with (i) higher proficiency populations than the CEFR A2 level, the pre-intermediate level in Fitzpatrick and Clenton (2017), and the beginner level in Horst and Collins (2006); and (ii) an extra productive writing vocabulary task (the LFP) to examine the extent to which CEFR B2 learners use of infrequent written productive vocabulary might change over time.

Accordingly, the research questions for the current chapter are:

RQ 1. To what extent do results from a longitudinal study suggest that three different productive vocabulary knowledge tasks might detect change for a group of B2 learners?

RQ 2. To what extent do results from a longitudinal study suggest that three productive vocabulary knowledge tasks might detect changes in aspects of writing for a group of B2 learners?

5.2 The study

In this chapter, I report on a study that explored productive vocabulary knowledge change over time as measured by Lex30, G_Lex, the PVLT, and the LFP.

5.2.1 Measures

The study employs the same productive vocabulary tasks from Chapters 3 and 4 (Lex30, G_Lex, and the PVLT, see Table 3.1) but with an additional productive vocabulary task (the LFP).

The Lexical Frequency Profile (the LFP). The lexical frequency profile (the LFP; Laufer & Nation, 1995; 2.1.1) is a measure of vocabulary use and has been used in a few studies (Fitzpatrick & Clenton, 2017; Laufer, 1994; Laufer & Nation, 1995). The LFP requires participants to write two paragraphs of about 300–350 words each. However, to allow longitudinal comparison across the four tasks, only one composition of 250 words was requested at each testing time. For ecological validity, I opted for a cap of 250 words in 40 minutes (rather than 350 words in 60 minutes as is conventionally required for the LFP task (Laufer & Nation, 1995, p. 314) to respond to the condition learners encounter in "high-stakes" writing tests (i.e., IELTS). Table 5.1 provides a summary of the differences between the LFP and the IELTS Writing Task 2.

Table 5.1

Differences between the LFP and IELTS Writing Task 2

	LFP	IELTS Writing Task 2
Number of compositions	Two	One*
Type of composition	Argumentative	Argumentative
Word count	300–350 words each	250 words
Time	60 minutes	40 minutes

* Laufer and Nation (1995) found that the LFP score was quite stable between compositions of different topics produced by the same students.

The two composition topics were adapted from IELTS (2019a) sample writing topics. The first LFP composition topic was "*The threat of nuclear weapons maintains world peace.*

Nuclear power provides cheap and clean energy. The benefits of nuclear technology far outweigh the disadvantages. To what extent do you agree or disagree?" The second LFP composition topic was "*The first car appeared on British roads in 1888. By the year 2000, there may be as many as 29 million vehicles on British roads. Alternative forms of transport should be encouraged and international laws introduced to control car ownership and use. To what extent do you agree or disagree*?" Following the original task scoring procedures, the completed LFP task papers were typed up and corpora were then compared with online frequency counts. To statistically compare learners' profiles across different tasks (Lex30, G_Lex, PVLT, and LFP) and as in Fitzpatrick and Clenton (2017), I used the same scoring protocol for all tests where the LFP score consists of a count of all but the highly frequent (i.e., non-1k) responses.

5.2.2 Participants

I assessed 50 (9 males and 41 females) second language (L2) undergraduate learners (of English) from a university in Southern France. The participants in this study were different from the participants in the last two studies. All participants reported considering French to be (one of) their native language(s). All participants were enrolled in their fourth semester of a six-semester degree in English studies program. Participants were on average 20.26-years-old (range: 18–27, SD = 1.44). Participants were at a higher proficiency level (CEFR B2) than the participants in Chapter 4 as reported by their classroom teacher.

Along with their coursework materials, participants were provided with self-study vocabulary activities. The self-study activities included a variety of academic vocabulary questions (e.g., true/false, multiple choice, sentence completion, matching, spelling). Participants were strongly encouraged to use the self-study vocabulary

words in their different class activities.

Participation in the study was voluntary with the right to opt out at any time. Personal data were protected as mainly the average mean test scores were reported in the study results with no personal information collected from any of the participants.

5.2.3 Methodology

Participants took two equivalent versions of the four productive vocabulary tasks: Lex30, G_Lex, PVLT, and LFP (Appendices 1–4). Testing was conducting at two points: at the beginning of their fourth semester of study and at the end of the semester three months later. To negate test effects, Tasks were presented in a different order at each of the two test times by the classroom teacher. Samples of participants' responses to the four measures are provided in Appendices 8-11.

As was the case in Chapters 3 and 4, individual responses to the productive vocabulary tasks were scored using the BNC-COCA online frequency corpora. A repeated measures ANOVA was conducted to explore the potential differences in the test mean percentage scores at the two time points.

As outlined in Chapter 3, percentage scores were used to address two main issues: (i) the different potential number of items for each task (because the PVLT elicits a potential maximum of 90 items, whereas both Lex30 and G_Lex elicit a potential maximum of 120 items, and the LFP elicit a potential maximum of 250); and (ii) the fact that different versions of the PVLT (Time 1 and Time 2), when processed with BNC-COCA, generated a different total number of scoring responses (the PVLT 1 (version A): BNC-COCA total score = 85; the PVLT 2 (version B): BNC-COCA total score = 79). The mean percentage scores represent (scoring items/possible scoring items x100) on all four tests at the two time points. Table 5.2 below presents the changes between the test mean percentage scores at the two testing times.

5.3 Results

Table 5.1 shows the different productive vocabulary task mean scores at the two different test times (Time 1, Time 2) for the CEFR B2 participants. A two-way repeated measures ANOVA was used to examine the effect of Task and Time on the test mean scores. While there was a significant effect for Task (F (3, 147) = 401.510, $p < .001$), no significant effect was revealed for Time (F (1, 49) = 1.028, $p = 0.315$). There was a statistically significant interaction between the effects of Task and Time (F (3, 147) = 14.847, $p < .001$). These results were further explored via post hoc comparison tests with Bonferroni correction. The post hoc tests showed that whereas scores on G_Lex significantly changed over time ($p <$ .001), no such significant change was recorded for Lex30 ($p = 1.000$), the PVLT *(*$p = .439$), or for the LFP ($p = .861$).

Table 5.2

Changes between the Test Mean Percentage Scores at the Two Testing Times for B2 Participants

$N = 50$	Mean Score			
	Lex30	G_Lex	PVLT	LFP
Time 1	33.45 (9.28)	15.73 (8.02)	53.48 (12.99)	16.71 (5.45)
Time 2	34.75 (10.12)	24.17 (7.76)	50.63 (15.07)	14.65 (6.34)
Changes between time points				
Time 2 − Time 1	1.30	8.44***	−2.85	−2.06
	(11.95)	(10.46)	(15.10)	(8.32)

Note: * $p < .05$, ** $p < .01$, *** $p < .001$ (Bonferroni correction). Numbers in parentheses are standard deviations.

5.4 Discussion

The introduction to this chapter outlined the need for a longitudinal investigation to explore the extent to which the four different tasks (Lex30, G_Lex, PVLT, and LFP) consistently detect the productive vocabulary knowledge change of higher-proficiency populations (CEFR B2) than those reported in Chapter 4. The research questions asked:
RQ1. To what extent do results from a longitudinal study suggest that three different productive vocabulary knowledge tasks might detect change for a group of A2 learners?
RQ2. To what extent do results from a longitudinal study suggest that three productive vocabulary knowledge tasks might detect changes in aspects of writing for a group of B2 learners?

A comparison of the findings in Chapter 5 with those of Chapter 4 might better respond to the research questions. Chapter 4 reported a longitudinal study in which I tested CEFR A2 level learners with Lex30, G_Lex, and the PVLT. Chapter 5 reported on a longitudinal study in which I tested CEFR B2 level learners with the same three tasks in addition to the LFP. The longitudinal data from the two different proficiencies allow me to make two important claims.

First, although the administered tasks claim to tap into the construct of productive vocabulary knowledge, they might detect different types of productive vocabulary knowledge once proficiency is factored into account. Chapter 4 (4.4) reports that Lex30 elicits a larger percentage of scoring items than the G_Lex and the PVLT tasks in both testing times for A2 learners (Time 1: Lex30 (12.20) > G_Lex (11.19) > PVLT (10.62) and Time 2: Lex30 (16.35) > G_Lex (13.30) > PVLT (11.32)). Chapter 5 (5.3) reports that the PVLT elicits a larger percentage of scoring items than the Lex30, G_Lex, and LFP tasks in both testing times for B2 learners (Time 1: PVLT (53.48) > Lex30 (33.45) > LFP (16.71) > G_Lex (15.73) and Time 2: PVLT (50.63) > Lex30 (34.75) > G_Lex (24.17) > LFP (14.65)). These

differences between the results in Chapters 4 and 5 suggest that factors other than task characteristics influence the test scores; for example, factors such as L2 proficiency (Milton, 2013; Walters, 2012) appear to play a role in determining the extent to which different productive vocabulary knowledge tasks detect change.

Second, the administered tasks uncovered inconsistent change over time for different proficiency levels (Table 5.3). For Chapter 4 A2 level learners, both Lex30 and G_Lex scores significantly increase over time but PVLT scores do not (Time 2 − Time 1: Lex30 (4.15), G_Lex (2.11), and PVLT (.7)). For Chapter 5 B2 level learners, G_Lex scores significantly increase over time but Lex30, the PVLT and the LFP scores do not (Time 2 - Time 1: G_Lex (8.44), Lex30 (1.30), LFP (−2.06), and PVLT (−2.85). These differences between the results in Chapters 4 and 5 suggest that L2 proficiency might play a role in determining the extent to which different productive vocabulary knowledge tasks detect consistent change.

Table 5.3

Significant Changes across Time as Measured by Lex30, G_Lex, PVLT, and LFP among Different Proficiency Levels

Task	Chapter 4 (A2 level)	Chapter 5 (B2 level)
Lex30	Significant change (4.15)	No significant change (1.30)
G_Lex*	Significant change (2.11)	Significant change (8.44)
PVLT	No significant change (.7)	No significant change (−2.85)
LFP	NA**	No significant change (−2.06)

* Only G_Lex showed significant change across the two proficiency levels.

** Horst and Collins (2006) report an inconsistent decrease in participants' infrequent written productive vocabulary knowledge when using the LFP in a longitudinal study with beginners L1 French.

The discussion above indicates that the administered productive vocabulary tasks might inconsistently detect different types of productive vocabulary knowledge at different proficiency levels. In general terms, tasks allude to a significant gain in participants' productive vocabulary knowledge but not all tasks (i.e., PVLT for CEFR A2 and B2 and Lex30, LFP for CEFR B2). In terms of relevant gains, the highest gain is seen in G_Lex scores followed by Lex30, LFP, and PVLT scores for the CEFR B2 participants, and for Lex30 scores followed by G_Lex then PVLT scores for the CEFR A2 participants. I will return to these discussion points of task consistency and gain in my discussion chapter.

5.5 Conclusion

This chapter longitudinally investigated the extent to which the four tasks of productive vocabulary knowledge (Lex30, G_Lex, the PVLT, and the LFP) detect productive vocabulary knowledge change over time for a group of B2 participants. The study was conducted over three months with testing at 0 and 3 months. Only G_Lex scores appear to indicate some degree of significant change; however, Lex30, the PVLT, and the LFP do not show significant change with this group of B2 participants.

To restate, Chapter 3 showed significant correlations of Lex30, G_Lex, and the PVLT tasks but did not indicate the extent to which these tasks might detect productive vocabulary knowledge over time. This, therefore, led to Chapter 4, which investigated the three tasks' ability to detect change among a group of A2 level learners. Chapter 5 extended Chapter 4's investigation of the tasks' ability to detect change but with a group of B2 level learners and an extra productive writing vocabulary task (the LFP). Both chapters (4 and 5) concluded that L2 proficiency (Milton, 2013; Walters, 2012) might influence the extent to which different productive vocabulary knowledge tasks detect change.

Having investigated writing, I turn now to speaking, the other productive skill. Moreover, in my final experimental chapter (Chapter 6), I wanted to return to investigating

A2 level students on the basis that the majority of L1 Japanese university students are at beginner to pre-intermediate L2 level (Barrow, Nakanishi, & Nishino, 1999; McLeana, Hogg, and Kramerc, 2014; Shillaw, 1995).

Earlier studies have addressed different aspects of speaking in relation to a single productive vocabulary task, e.g., Lex30 (Clenton et al., 2021) and the PVLT (De Jong et al., 2012), suggesting that aspects of productive vocabulary knowledge and the ability to showcase such knowledge might vary according to proficiency. In Clenton et al. (2021), participants were of pre-intermediate level and in De Jong et al. (2012), participants were of intermediate to advanced level proficiency level. The final experimental chapter in this thesis examines the extent to which productive vocabulary knowledge detected by three productive vocabulary tasks (Lex30, G_Lex, and the PVLT) might consistently relate to aspects of the multifaceted construct of fluency.

Chapter 6 Investigating the Extent to which Three Productive Vocabulary Knowledge Tasks Consistently Detect Aspects of Oral Fluency Change among an A2 English User Group

6.1 Introduction

A comparison of the findings in Chapter 5 with those of Chapter 4 (Section 5.4) indicated that L2 proficiency (i.e., the ability to produce words in context and contextually demanding tasks) may play a role in determining the gain of vocabulary knowledge elicited by a given task. In Chapter 4, the highest gain was seen in Lex30 scores followed by G_Lex, and then the PVLT scores for the CEFR A2 participants. In Chapter 5, the highest gain was seen in G_Lex scores followed by Lex30, the LFP, and then the PVLT scores for the CEFR B2 participants. The different gains among the two proficiency levels might support the claim that the extent to which productive vocabulary knowledge tasks detect change might be proficiency-based.

In the current chapter, I continue the theme for the earlier chapters (3-5) by using the same three productive vocabulary tasks (G_Lex, Lex30, and PVLT). I return to investigating A2 level L1 Japanese university students on the basis that a large population of L1 Japanese university students are at beginner to pre-intermediate L2 level, with an average vocabulary size of between 2,300 (Barrow, Nakanishi, & Nishino, 1999; Shillaw, 1995) to 3,715 words (McLean, Hogg, & Kramer, 2014). Having investigated writing for B2 level in Chapter 5, I now turn to “speaking.”

One of the main indicators of general speaking ability is fluency, as reflected in the rating descriptions of standardized English tests (i.e., IELTS). While the broad sense of fluency appears to relate to overall or global proficiency, the narrow sense of fluency (for diagnostic purposes) “refer(s) to one, presumably isolatable, component of oral proficiency” (Lennon, 1990, p. 389). Within narrow fluency, fluency is often measured as a component of

speech with multiple aspects, referring to quick and perhaps "smooth delivery of speech without (filled) pauses, repetitions, and repairs" (De Jong et al., 2015, p. 235).

Earlier papers on fluency have explored fluency in relation to a single productive vocabulary task (e.g., the PVLT) (De Jong et al., 2012) and Lex30 (Clenton et al., 2021, Table 6.1). De Jong et al. (2012) administered a cross-sectional study of 179 intermediate to advanced level adult learners of Dutch with eight computer-administered, semi-spontaneous speaking tasks designed to elicit aspects of L2 fluency in relation to "L2 linguistic knowledge and processing skills" ranging in terms of complexity, formality, and discourse type, and a newly constructed Dutch version of the PVLT.

Clenton et al. (2021) administered a cross-sectional study of 30 pre-intermediate undergraduate adult L1 learners of English with three speaking tasks (from De Jong et al. (2012)), a productive vocabulary task (Lex30), and a receptive vocabulary task (X_Lex). Both studies were (i) cross-sectional, (ii) used experimental speaking tasks, and (iii) confirmed significant correlations between vocabulary knowledge and aspects of fluency (i.e., significant negative correlation between number of silent pauses and PVLT ($r = -.39$, $p < .05$) for De Jong et al.'s participants and Lex30 ($r = -.39$, $p < .05$) for Clenton et al.'s participants.

The current chapter builds on De Jong et al. (2012) and Clenton et al.'s (2021) suggestion of follow up studies to investigate consistent relations between additional and different vocabulary measures and aspects of fluency. The current study, therefore, is (i) longitudinal, (ii) employs multiple productive vocabulary tasks (G_Lex, Lex30, and PVLT), and (iii) employs a full IELTS speaking test (for ecological validity, IELTS is a common high-stakes test in Japan).

Table 6.1

Speaking in Relation to a Single Productive Vocabulary Task

Study	Participants	Tasks	Results	Findings
De Jong et al. (2012)	179 intermediate to advanced level learners of Dutch	Eight computer-administered, semi-spontaneous speaking tasks Dutch version of the PVLT	Both studies confirmed significant correlations between vocabulary knowledge and aspects of fluency (i.e., significant negative correlation between number of silent pauses and PVLT ($r = -.39$, $p < .05$) for De Jong et al.'s participants and Lex30 ($r = -.39$, $p < .05$) for Clenton et al.'s participants)	Both studies suggest that vocabulary knowledge and fluent speech may to some extent be proficiency-dependent
Clenton et al. (2021)	30 pre-intermediate level learners of English	Three speaking tasks from De Jong et al. (2013) Lex30 X_Lex		

Accordingly, the research question for the current chapter is:

RQ. To what extent do results from a longitudinal study suggest that three different productive vocabulary knowledge tasks might detect changes in aspects of speaking fluency for a group of A2 learners?

6.2 The study

In this chapter, I report on a study that explores the extent to which three different productive vocabulary knowledge tasks detect change over time in relation to different speaking fluency measures.

6.2.1 Measures

The study employs the same productive vocabulary tasks from Chapters 3-5 (Lex30, G_Lex, and PVLT, see Table 3.1) but with a full IELTS speaking test as measured by three speaking fluency measures (phonation rate (De Jong, 2012), speech rate (Segalowitz, 2010), and articulation rate (De Jong, 2012).

IELTS Speaking Test. The IELTS speaking test assesses the use of spoken English with a rubric of four main band descriptors: fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation. The test lasts between 11 and 14 minutes during which test-takers discuss different topics with an IELTS examiner. The test contains three parts:

Part 1: Introduction and Interview. This part uses a question–response format to assess the test-taker's ability to respond to everyday topics. The examiner asks the test-taker general questions about themselves (e.g., current living places, work, study, etc.) and a range of familiar topics (e.g., favorite music, food, weather, movies, etc.). The examiner might ask some follow-up questions, e.g., "why" or "why not" in order to extend the discussion and elicit more speaking from the test-taker. This part lasts between 4 and 5 minutes.

Part 2: Individual Long Turn. This part uses a card–response format to assess the test-taker's ability to speak for a longer time on a specific topic with appropriate language and ideas. The examiner gives the test-taker a task card with a speaking prompt and asks them to respond to it. The topics usually require the test-taker to

describe or state their opinion on certain topics (e.g., “describe something you own that is very important to you” (IELTS, 2019a). Test-takers should have a minute for preparation, followed by up to two minutes of speaking. The examiner might ask a couple of questions on the same topic.

Part 3: Two-way Discussion. This part uses a follow-up questions format to assess the test-taker’s ability to discuss abstract issues and ideas. The examiner asks more questions related to the topic in Part 2. As an example, in relation to the earlier speaking topic in part 2, a follow-up question from the examiner might be: “What makes things important for you?” This part of the test takes about 4–5 minutes to complete.

6.2.2 Participants

I assessed 40 (16 males and 24 female) second language (L2) undergraduate learners (of English) from a university in western Japan, all with L1 Japanese. The participants in this study were different from the participants in the last three studies. All participants were aged 18 years and in their first year at university. Participants’ TOEIC Bridge score ranged between 128 and 138, indicating a proficiency level of A2 according to the Common European Framework of Reference (CEFR) (ETS, 2019). All participants were placed in level 1 (lowest level of three) compulsory skills-based academic English classes (Listening II and Listening V) for 90 minutes a week.

Along with their coursework materials, participants were provided with self-study vocabulary activities. The self-study activities included a variety of academic vocabulary questions (e.g., true/false, multiple choice, sentence completion, matching, spelling). Participants were strongly encouraged to use the self-study vocabulary words in their different class activities.

Participation in the study was voluntary with the right to opt out at any time. Personal data were protected as mainly the average mean test scores were reported in the study results

with no personal information collected from any of the participants.

6.2.3 Methodology

Participants took two equivalent versions of the three productive vocabulary tasks: Lex30, G_Lex, and the PVLT (Appendices 2-4) along with two full equivalent versions of IELTS speaking test. I conducted testing at two test points: at the beginning of their first semester of study and at the end of the same semester three months later. To negate test effects, I presented the productive vocabulary tasks in a different order at each of the two test times.

As was the case in Chapters 3-5, individual responses to the productive vocabulary tasks (Lex30, G_Lex, and the PVLT) were scored using the BNC-COCA online frequency corpora. As outlined in Section 3.2.3, percentage scores were used to address two main issues: (i) the different potential number of items for each task (because the PVLT elicits a potential maximum of 90 items, whereas both Lex30 and G_Lex elicit a potential maximum of 120 items) and (ii) the fact that different versions of the PVLT (Time 1 and Time 2), when processed with BNC-COCA, generate a different total number of scoring responses (the PVLT 1 (version A): BNC-COCA total score = 85 and the PVLT 2 (version B): BNC-COCA total score = 79). The mean percentage scores represent (scoring items/possible scoring items x100) on all four tests at the two time points.

The IELTS speaking tests were administered according to their original procedures by a certified IELTS examiner. To measure aspects of fluency, (i) the speaking tests were recorded and then transcribed; (ii) word syllables were counted for all participants; (iii) participants' speaking time was counted and abstracted from the examiner's speaking time; and (iv) participants' audio recordings were analyzed using PRAAT (Boersma & Weenink, 2005). Silent pauses were measured manually at 350 ms (as Clenton et al., 2021 and De Jong et al., 2012). All sounds (e.g., *ehh, uhh,* and *umm)* were recorded as filled pauses. All

measures were collated over the three parts of the IELTS speaking test. Subsequently, three fluency measures were calculated per second of speaking: (i) phonation rate (De Jong, 2012), i.e., speaking time with no pauses/speaking time with pauses; (ii) speech rate (Segalowitz, 2010), i.e., syllable count/speaking time with pauses; and (iii) articulation rate (De Jong, 2012), i.e., syllable count/speaking time with no pauses.

6.3 Results

6.3.1 Time Effect (Time 1 → Time 2) on Different Fluency Measures

Table 6.2 reports participants' mean scores for fluency as measured by phonation rate, speech rate, and articulation rate for the three different speaking tasks (Part 1, Part 2, and Part 3) at Time 1 (Pre-time) and Time 2 (Post-time), and their changes over time. A two-way repeated measures ANOVA was used to examine the effect of Time on the mean scores for fluency measures at different speaking tasks. There was a significant main effect for Fluency (F (3.3, 129.1) = 364.83, $p < .001$) and statistically non-significant effect for Time (F (1, 39) = 1.188, $p = .282$). Interaction between the effects of Time and Fluency is also non-significant (F (2.78, 108.57) = 1.146, $p = .332$).

Post hoc comparison tests with Bonferroni correction were then performed to compare the mean differences across Time for each fluency measure. The post hoc tests showed (i) no significant change in phonation rate from Time 1 to Time 2 for all three speaking tasks (Part 1, Part 2, and Part 3) and (ii) a significant decrease in speech rate on Time 2 (Post) in Part 1 speaking task while there is a significant increase in speech rate on Time 2 (Post) in Part 3. For Part 2, no significant change is observed in speech rate from Time 1 to Time 2 and (iii) there is no significant change in articulation rate from Time 1 to Time 2 for all three speaking tasks (Part 1, Part 2, and Part 3).

Table 6.2

Fluency Measure Mean Scores for Three Different Speaking Tasks

Speaking Task	Fluency Measure	Mean (std. dev) at Time 1	Mean (std. dev) at Time 2	Changes in Mean from Time 1 to Time 2
Part 1	Phonation rate	0.41 (0.132)	0.40 (0.107)	−.010 (0.133)
	Speech rate	1.31 (0.439)	1.13 (0.256)	−.173* (0.436)
	Articulation rate	3.33 (1.218)	2.95 (0.808)	−.383(1.485)
Part 2	Phonation rate	0.43 (0.136)	0.43 (0.106)	−.001 (0.176)
	Speech rate	1.18 (0.361)	1.13 (0.423)	−.046 (0.445)
	Articulation rate	3.06 (2.209)	2.82 (1.296)	−.238 (2.730)
Part 3	Phonation rate	0.33 (0.100)	0.37 (0.112)	.038 (0.134)
	Speech rate	0.95 (0.287)	1.06 (0.300)	.102* (0.258)
	Articulation rate	2.96 (0.682)	2.99 (0.979)	−.032 (1.219)

Note: * $p < .05$, ** $p < .01$, *** $p < .001$ (Bonferroni correction). Numbers in parentheses are standard deviations.

6.3.2 Time Effect (Time 1 → Time 2) on Different Vocabulary Test Scores

Table 6.3 shows participants' vocabulary test percentage mean scores for Lex30, G_Lex, and PVLT at Pre-Time 1 and Post-Time 2, and their changes over time. A two-way repeated measures ANOVA was used to examine the effect of Time on Test scores. There was significant main effect for Test (F (2, 78) = 20.134, $p < .001$) and non-significant main effect of Time (F (1, 39) = 2.344, $p = .134$). Interaction between the effects of Time and Test is significant (F (2, 78) = 32.795, $p < .001$).

Post hoc comparison tests with Bonferroni correction were then performed to compare the mean differences across Time and Test. The post hoc tests showed (i) no significant change in Lex30 score from Time 1 to Time 2, (ii) a significant improvement in G_Lex Test score from Time 1 to Time 2, and (iii) a significant decrease in PVLT test scores from Time 1 to Time 2.

Table 6.3

Vocabulary Scores at Time 1 and Time 2

Vocabulary Test	Mean (std. dev) at Time 1	Mean (std. dev) at Time 2	Changes in Mean from Time 1 to Time 2
Lex 30	13.69 (5.982)	15.08 (7.190)	1.396 (6.823)
G_Lex	7.06 (3.418)	12.17 (5.498)	5.104* (4.867)
PVLT	14.94 (5.269)	11.01 (4.570)	−3.929* (4.170)

Note: * $p < .05$, ** $p < .01$, *** $p < .001$ (Bonferroni correction). Numbers in parentheses are standard deviations.

6.3.3 Relationship between Vocabulary and Fluency at Time 1_Part 1

Table 6.4 shows the correlation between the vocabulary test scores and the fluency measures for Part 1 of the speaking task at Time 1. There is a moderate level of correlation between vocabulary test scores and fluency measures in the Part 1 speaking task. There is significant correlation between G_Lex test score and phonation rate ($r(38) = .460$, $p = .003$). There is significant negative correlation between PVLT test score and articulation rate ($r(38) = .340$, $p = .032$). Within the vocabulary test scores and fluency measures, there are moderate levels of significant correlation. Correlation between Lex30 and G_Lex is $r(38) = .544$, $p < .001$. Significant correlation between phonation rate and speech rate is $r(38) = .458$, $p = .003$ and speech rate and articulation rate is $r(38) = .544$, $p < .001$ while the phonation rate and articulation rate show moderate negative correlation ($r(38) = -.446$, $p = .004$).

Table 6.4

Correlations between Vocabulary Test and Fluency Measures at Time 1_Part 1

	Lex30	G_Lex	PVLT	Phonation Rate	Speech Rate	Articulation Rate
Lex30	1	.544**	.128	.107	.162	.067
G_Lex		1	.117	.460**	.251	−.122
PVLT			1	.162	−.140	−.340*
Phonation rate				1	.458**	−.446**
Speech rate					1	.544**
Articulation rate						1

Note: * $p < .05$, ** $p < .01$, *** $p < .001$ (Bonferroni correction).

6.3.4 Relationship between Vocabulary and Fluency at Time 2_ Part 1

Table 6.5 shows the correlation between the vocabulary test scores and the fluency measures for Part 1 of the speaking task at Time 2. There is no significant correlation observed between any of vocabulary test scores and fluency measures in the Part 1 speaking task at Time 2. Within the vocabulary test scores and fluency measures, there are moderate levels of significant correlation. Correlation between Lex30 and G_Lex is $r(38) = .664$, $p < .001$, between Lex30 and PVLT test scores is $r(38) = .445$, $p = .004$, and between G_Lex and PVLT is $r(38) = .407$, $p = .009$). Significant correlation between speech rate and articulation rate is $r(38) = .492$, ($p = .001$) while the Phonation rate and Articulation rate show moderate negative correlation ($r(38) = -.617$, $p < .001$). Correlation between phonation and speech rates is non-significant ($p = .093$).

Table 6.5

Correlations between Vocabulary Test and Fluency Measures Time 2_Part 1

	Lex30	G_Lex	PVLT	Phonation rate	Speech rate	Articulation rate
Lex30	1	.664**	.445**	.155	−.022	−.170
G_Lex		1	.407**	.306	.042	−.131
PVLT			1	.161	.114	−.058
Phonation rate				1	.269	−.617**
Speech rate					1	.492**
Articulation rate						1

Note: * $p < .05$, ** $p < .01$, *** $p < .001$ (Bonferroni correction).

6.3.5 Time Effect (Time 1 → Time 2) on Change in Relationship between Vocabulary Test Scores and Fluency Measures_ Part 1

Table 6.6 shows correlations between the vocabulary test scores and the fluency measures for Pre (Time 1) and Post (Time 2) levels. The data show (i) an insignificant correlation between Lex30 test and fluency measures on both Pre- and Post–Time, (ii) a significant weak correlation between G_Lex and phonation rate at Time 1, which becomes non-significant at Time 2, and (iii) significant negative correlation between PVLT and articulation rate at Time 1, which becomes non-significant at Time 2.

Table 6.6

Correlations between Vocabulary Test and Fluency Measures at Time 1 and Time 2_Part 1

	Phonation Rate		Speech Rate		Articulation Rate	
	Time 1	Time 2	Time 1	Time 2	Time 1	Time 2
Lex30	.107	.155	.162	−.022	.067	−.170
G_Lex	.460**	.306	.251	.042	−.122	−.131
PVLT	.162	.161	−.140	.114	−.340*	-.058

Note: * $p < .05$, ** $p < .01$, *** $p < .001$ (Bonferroni correction).

6.3.6 Relationship between Vocabulary and Fluency at Time 1_Part 2

Table 6.7 shows correlations of the vocabulary test scores and the fluency measures for Part 2 of the speaking task at Time 1. There is no significant correlation observed between any of vocabulary test scores and fluency measures in Part 2 speaking task at Time 1. Within the vocabulary test scores and fluency measures there are moderate levels of significant correlation. Correlation between Lex30 and G_Lex is $r(38) = .544$, $p < .001$. Significant correlation between phonation rate and speech rate is $r(38) = .613$, $p < .001$ and for speech rate and articulation rate is $r(38) = .394$, $p = .012$ while the phonation rate and articulation rate show moderate negative correlation ($r(38) = -.465$, $p = .003$).

Table 6.7

Correlations between Vocabulary Test and Fluency Measures at Time 1_Part 2

				Phonation		Articulation
	Lex30	G_Lex	PVLT	Rate	Speech Rate	Rate
Lex30	1	.544**	.128	−.219	−.097	.199
G_Lex		1	.117	.016	.130	.179
PVLT			1	.058	.159	.172
Phonation rate				1	.613*	−.465**
Speech rate					1	.394*
Articulation rate						1

Note: * $p < .05$, ** $p < .01$, *** $p < .001$ (Bonferroni correction).

6.3.7 Relationship between Vocabulary and Fluency at Time 2_Part 2

Table 6.8 shows correlations of the vocabulary test scores and the fluency measures for Part 2 of the speaking task at Time 2. There is significant moderately negative correlation observed between any of the PVLT test scores and articulation rate in the Part 2 speaking task at Time 2 ($r(38) = -.421$, $p = .007$). Within the vocabulary test scores and fluency measures there are moderate levels of significant correlation. Correlation between Lex30 and G_Lex is $r(38) = .664$, $p < .001$ and between Lex30 and PVLT test scores $r(38) = .445$, ($p = .004$), and between G_Lex and PVLT $r(38) = .407$, ($p = .009$). The significant correlation between phonation rate and speech rate is $r(38) = .431$, $p = .005$, and speech rate and articulation rate is $r(38) = .543$, $p < .001$ while that between phonation rate and articulation rate has a moderate negative correlation ($r(38) = -.438$, $p = .005$).

Table 6.8

Correlations between Vocabulary Test and Fluency Measures at Time 2_Part 2

	Lex30	G_Lex	PVLT	Phonation Rate	Speech Rate	Articulation Rate
Lex30	1	.664**	.445**	.193	−.192	−.201
G_Lex		1	.407**	.058	−.224	−.243
PVLT			1	.257	−.283	−.421**
Phonation rate				1	.431**	−.438**
Speech rate					1	.543**
Articulation rate						1

Note: * $p < .05$, ** $p < .01$, *** $p < .001$ (Bonferroni correction).

6.3.8 Time Effect (Time 1 → Time 2) on Change in Relationship between Vocabulary Test Scores and Fluency Measures_Part 2

Table 6.9 shows correlations between the vocabulary test scores and the fluency measures for Pre (Time 1) and Post (Time 2). It can be concluded from the Part 2 speaking task that (i) there is a non-significant correlation between the Lex30 test and fluency measures on both Pre- and Post- Time, (ii) there is a non-significant correlation between G_Lex test and fluency measures on both Pre- and Post-Time, and (iii) there is a non-significant positive correlation between PVLT and articulation rate at Time 1, which emerges as a moderate negative significant correlation at Time 2.

Table 6.9

Correlations between Vocabulary Test and Fluency Measures at Time 1 and Time 2_Part 2

	Phonation Rate		Speech Rate		Articulation Rate	
	Time 1	Time 2	Time 1	Time 2	Time 1	Time 2
Lex30	−.219	.193	−.097	−.192	.199	−.201
G_Lex	.016	.058	.130	−.224	.179	−.243
PVLT	.058	.257	.159	−.283	.172	−.421**

Note: * $p < .05$, ** $p < .01$, *** $p < .001$ (Bonferroni correction).

6.3.9 Relationship between Vocabulary and Fluency at Time 1_Part 3

Table 6.10 shows correlations between the vocabulary test scores and the fluency measures for Part 3 of the speaking task at Time 1. There is a moderate level of correlation between vocabulary test scores and fluency measures. There is significant correlation between G_Lex test score and phonation rate ($r(38) = .490$, $p = .001$). There is significant correlation between the G_Lex test score and speech rate ($r(38) = .372$, $p = .018$). Within the vocabulary test scores and fluency measures there are moderate levels of significant correlation. Correlation between Lex30 and G_Lex is $r(38) = .544$, $p < .001$. Correlation between phonation rate and speech rate is $r(38) = .660$, $p < .001$ while the phonation rate and articulation rate show moderate negative correlation ($r(38) = -.482$, $p = .002$).

Table 6.10

Correlations between Vocabulary Test and Fluency Measures at Time 1_Part 3

	Lex30	G_Lex	PVLT	Phonation Rate	Speech Rate	Articulation Rate
Lex30	1	.544**	.128	.110	.272	.153
G_Lex		1	.117	.490**	.372*	−.177
PVLT			1	−.066	−.085	.060
Phonation rate				1	.660**	−.482**
Speech rate					1	.306
Articulation rate						1

Note: * $p < .05$, ** $p < .01$, *** $p < .001$ (Bonferroni correction).

6.3.10 Relationship between Vocabulary and Fluency at Time 2_Part 3

Table 6.11 shows the correlations of the vocabulary test scores with the fluency measures for Part 3 of the speaking task at Time 2. There is no significant correlation observed between any of vocabulary test scores and fluency measures in Part 3 of the speaking task at Time 2. Within the vocabulary test scores and fluency measures there are moderate levels of significant correlation. Correlation between Lex30 and G_Lex is $r(38) = .664$, $p < .001$, between Lex30 and PVLT test scores $r(38) = .445$, $p = .004$, and between G_Lex and PVLT $r(38) = .407$, $p = .009$). Significant correlation between phonation rate and speech rate is $r(38) = .567$, $p < .001$, and between speech rate and articulation rate $r(38) = .380$, $p = .016$ while phonation rate and articulation rate has moderate negative correlation ($r(38) = -.496$, $p = .001$).

Table 6.11

Correlations between Vocabulary Test and Fluency Measures at Time 2_Part 3

	Lex30	G_Lex	PVLT	Phonation Rate	Speech Rate	Articulation Rate
Lex30	1	.664**	.445**	.006	.108	.179
G_Lex		1	.407**	.107	.191	.074
PVLT			1	.220	.035	−.134
Phonation rate				1	.567**	−.496**
Speech rate					1	.380*
Articulation rate						1

Note: * $p < .05$, ** $p < .01$, *** $p < .001$ (Bonferroni correction).

6.3.11 Time Effect (Time 1 → Time 2) on Change in Relationship between Vocabulary Test Scores and Fluency Measures_Part 3

Table 6.12 shows correlations between the vocabulary test scores and the fluency measures for Pre (Time 1) and Post (Time 2). It can be concluded at Part 3 of the speaking task that (i) there is a non-significant correlation between Lex30 test and fluency measures on both Pre- and Post-Time, (ii) there is a non-significant correlation between PVLT test and fluency measures on both Pre- and Post-Time, and (iii) there was a significant weak correlation between G_Lex and phonation Rate and speech rate at Time 1, which becomes non-significant at Time 2.

Table 6.12

Correlations between Vocabulary Test and Fluency Measures at Time 1 and Time 2_Part 3

	Phonation Rate		Speech Rate		Articulation Rate	
	Time 1	Time 2	Time 1	Time 2	Time 1	Time 2
Lex30	.110	.006	.272	.108	.153	.179
G_Lex	.490**	.107	.372*	.191	−.177	.074
PVLT	−.066	.220	−.085	.035	.060	−.134

Note: * $p < .05$, ** $p < .01$, *** $p < .001$ (Bonferroni correction).

6.4 Discussion

The introduction to this chapter outlined the need for a longitudinal study to explore potential relationships between productive vocabulary knowledge task scores and aspects of fluency. The study reported in Chapter 6 employed multiple productive vocabulary tasks (Lex30, G_Lex, and the PVLT) and an IELTS speaking test for a group of CEFR A2 participants. The research question asked: To what extent do results from a longitudinal study suggest that three different productive vocabulary

knowledge tasks might detect changes in aspects of speaking fluency for a group of A2 learners?

The participants took two equivalent versions of three productive vocabulary tasks (Lex30, G_Lex, and the PVLT), and a full speaking IELTS test. Testing was conducted at two test points: at the beginning of their first semester of study and at the end of the same semester, three months later. I included an analysis of the vocabulary task scores and three fluency measures (phonation rate, speech rate, and articulation rate) for the two testing times.

6.4.1 Relationship between Productive Vocabulary Scores and Aspects of Fluency

In broad terms, the results are to some extent consistent with earlier fluency studies (e.g., Clenton et al., 2021; De Jong et al., 2012) in that a vocabulary score might significantly correlate with aspects of fluency. In the current study, and as illustrated in Table 6.13, both G_Lex and the PVLT correlate significantly with aspects of fluency. G_Lex correlates significantly with fluency in the first testing time, i.e., phonation rate in Part 1 ($r = .460$, $p < .01$), phonation rate in Part 3 ($r = .490$, $p < .01$), and speech rate in Part 3 ($r = .372$, $p < .05$). The PVLT results correlate significantly (and negatively) with fluency in the first testing time, i.e., the articulation rate in Part 1 ($r = -.340$, $p < .05$) and in the second testing time with the articulation rate in Part 2 ($r = -.421$, $p < .01$). Such findings support the claim that vocabulary score might significantly relate to aspects of fluency at a pre-intermediate level of proficiency for the specific participants examined in the current study.

Table 6.13

Correlations between Vocabulary Tasks and Fluency Measures at Time 1 and Time 2

Tasks		Part 1			Part 2			Part 3		
	Time	Phonation Rate	Speech Rate	Articulation Rate	Phonation Rate	Speech Rate	Articulation Rate	Phonation Rate	Speech Rate	Articulation Rate
Lex30	T1	.107	.162	.067	−.219	−.097	.199	.110	.272	.153
G_Lex		.460**	.251	−.122	.016	.130	.179	.490**	.372*	−.177
PVLT		.162	−.140	-.340*	.058	.130	.172	−.066	−.085	.060
Lex30	T2	.155	−.022	−.170	.193	−.192	−.201	.006	.108	.179
G_Lex		.306	.042	−.131	.058	−.224	−.243	.107	.191	.074
PVLT		.161	.114	−.058	.257	-.283	-.421**	.220	.035	−.134

Note: * $p < .05$, ** $p < .01$, *** $p < .001$ (Bonferroni correction).

On the other hand, and unlike Clenton et al. (2021), the current study did not find any significant correlation between Lex30 task scores and the three fluency measures at either of the two testing times. I assume that such difference in the findings between both studies (current study and Clenton et al., 2021) might relate to: (i) the different speaking tasks, i.e., a full IELTS speaking test for the current study (used for ecological validity because IELTS is a common high-stakes test in Japan) and the three experimental speaking tasks from De Jong et al. (2012) for Clenton et al. (2021); (ii) the different vocabulary corpora, e.g., BNC-COCA (Nation, 2017) for the current study and the "Academic Spoken Word List" (ASWL; Dang, Coxhead, and Webb, 2017) for Clenton et al. (2021); and (iii) the different procedures used, e.g., the Lex30-score was processed based on percentage (scoring items/possible score x100) for the current study and based on the raw score (non-1k) for Clenton et al. (2021).

6.4.2 Consistent Relationships between Productive Vocabulary Scores and Aspects of Fluency

The results from the longitudinal data show that aspects of fluency do not appear to relate consistently to productive vocabulary knowledge task scores for this group of A2 participants. While G_Lex and the PVLT correlate significantly with aspects of fluency as mentioned in Section 6.4.1, only the PVLT shows a relationship across the two testing times. The PVLT score is negatively significant with the articulation rate in both testing times (Time 1 Part 1 ($r = -.340$, $p < .05$) and Time 2 Part 2 ($r = -.421$, $p < .01$)).

This finding might first relate to the time constraint. In general terms, the articulation process might possess a heavier cognitive load on L2 speakers (compared with the PVLT as a written task) under a strict time constraint (De Jong, 2016). The time allowed for the speaking tasks were 4−5 minutes for Part 1 and 2 minutes for Part 2, whereas 30 minutes were needed for the PVLT.

Second, the finding might relate to the task constraint or, in other words, how both the PVLT and the speaking tasks are being administered and processed. Unlike Lex30 and G_Lex, the PVLT was mainly designed to elicit predetermined infrequent (non-1k) words. The PVLT score is a count of all written correct answers. Unlike the PVLT, the articulation rate of the speaking tasks considers all spoken words (including the 1k list). The articulation rate is an account of all spoken words' syllables/total speaking time. Referring to both tests' (the PVLT and the speaking tasks) characteristics and score processing, participants had to use mainly infrequent words for the PVLT but not for the speaking tasks. I assume, therefore, that participants' responses were within the non-1k for the PVLT and, conversely, mainly within the 1k for the speaking tasks. An analysis of the 1k percentage in the participants' spoken words elicited by the three IELTS speaking tasks with BNC-COCA has supported my assumption: for Time 1, the 1k is 81.81% and for Time 2, the 1k is 79.74%.

6.5 Conclusion

This chapter explored the extent to which three different productive vocabulary tasks might detect aspects of fluency for a group of A2 participants. Participants (n = 45) took two equivalent versions of the three productive vocabulary tasks and full speaking IELTS test. Testing was conducted twice, at 0 (pre) and 3 months (post). The three productive vocabulary tasks were processed with BNC-COCA. The full IELTS speaking test was processed by PRAAT (Boersma & Weenink, 2005). The three measures of fluency were calculated per second of speaking: (i) phonation rate (De Jong, 2012), i.e., speaking time with no pauses/speaking time with pauses; (ii) speech rate (Segalowitz, 2010), i.e., syllable count/speaking time with pauses; and (iii) articulation rate (De Jong, 2012), i.e., syllable count/speaking time with no pauses.

The results revealed that vocabulary task scores might not consistently relate to aspects of fluency. As illustrated in Table 6.13, both G_Lex and the PVLT show moderate but significant correlations with aspects of fluency. Only the PVLT (when compared with Lex30 and G_Lex) demonstrated a significant relationship with aspects of fluency (i.e., the articulation rate) in both testing times (Time 1 Part 1 and Time 2 Part 2). Lex30 did not show any significant relationship at either testing time.

Chapter 7 Discussion

Introduction

To restate, this thesis aims to (i) investigate the extent to which productive vocabulary knowledge tasks can detect changes over a short time for two different proficiency groups and (ii) how such a change might relate to aspects of speaking and writing abilities. The experimental chapters reported four main findings: (i) productive vocabulary knowledge is task-dependent for groups of A2 and B2 proficiency participants; (ii) productive vocabulary knowledge change varies according to task and proficiency for groups of A2 and B2 proficiency participants; (iii) infrequent words use in writing consistently relates to productive vocabulary knowledge task (the PVLT) score for a group of B2 proficiency participants; and (iv) aspects of L2 oral fluency inconsistently relate to productive vocabulary knowledge task scores for a group of A2 proficiency participants.

In light of my experimental chapters' findings, the following discussion is divided into two main sections. The first section (7.1) discusses three issues that relate to productive vocabulary tasks sensitivity to detect change. I first explore the extent to which the productive vocabulary tasks' sensitivity to detect change might relate to task characteristics. Second, I discuss the extent to which the productive vocabulary tasks' sensitivity to detect change might relate to a proposed theory of L2 word development by Jiang (2000; 2002; 2004). Third, I discuss the productive vocabulary tasks' sensitivity to detect consistent change.

The second section (7.2) discusses three issues that relate more broadly to the productive vocabulary tasks' sensitivity to detect changes in infrequent words use in IELTS writing and speaking. First, I discuss how productive vocabulary tasks might detect change in infrequent words use in the IELTS writing task (for B2 participants). Second, I discuss how productive vocabulary tasks might detect infrequent words change in IELTS speaking tasks

(for A2 participants). Third, I explore potential pedagogical implications related to IELTS writing and speaking courses.

7.1 Productive Vocabulary Tasks' Sensitivity to Detect Change

As outlined in earlier research (i.e., Fitzpatrick & Clenton, 2017), the construct of productive vocabulary knowledge is far from straightforward. The array of overlapping interpretations as to what is actually covered by the construct has led to the development of a myriad of tasks whose aim is the same, namely, to measure productive vocabulary knowledge. However, upon closer examination, as demonstrated in the Section 3.5 discussion of Lex30, G_Lex and the PVLT task characteristics, tasks appear to make different assumptions about the underlying aspects of vocabulary knowledge of which the construct consists.

7.1.1 Productive Vocabulary Tasks' Sensitivity to Detect Change According to Task Characteristics

The three productive vocabulary knowledge tasks (Lex30, G_Lex, and PVLT), as outlined in Section 3.5, represent three contrasting but to some extent overlapping scholarly interpretations of the productive vocabulary construct. All three tasks activate user knowledge/ability (Meara, 1997); however, differences in the task characteristics (e.g., the extent to which tasks encourage the task-taker to consider context) (see Chapter 3, Table 3.1, and Figure 3.4 for a description of the three tasks' characteristics) might impact (i) the quantity of vocabulary knowledge a learner can showcase and (ii) the type of words being elicited by each task.

More specifically, as sampled in Figures 7.1 and 7.2, Lex30 elicits single-word items in response to its single-word cues. I suggest that Lex30 likely and potentially elicits a greater number of items than G_Lex and the PVLT because Lex30 appears to elicit collocations that respond to the activated cues (e.g., for the cue "away" the A2 level responds with collates

"*go*, *move*, and *run"* while the B2 level responds with collates "*go*, *fly*, and *run"*), which is likely less demanding than to provide responses potentially requiring attention to context (both in terms of part of speech and in terms of appropriate meaning) as is seemingly required for the G_Lex and PVLT tasks. I expect, therefore, Lex30 to be the most sensitive to detect change because the task demands are the least for Lex30 in comparison with both G_Lex and the PVLT.

G_Lex uses sentence cues to elicit single-word responses. I suggest that G_Lex likely and potentially elicits a smaller number of items than the Lex30 task because G_Lex appears to elicit responses that appear to require attention to context in terms of semantic appropriateness and grammatical accuracy from the sentence elicitation cues (e.g., for the missing gap in "He liked to ……. in his free time," A2 level participants respond with "*cook*, *sleep*, *walk*, *run*, and *read"* while B2 level participants respond with "*play*, *sleep*, *walk*, *go out*, and *drink,"* which is likely more demanding than to provide collocations as in Lex30. I expect, therefore, G_Lex to be less sensitive than Lex30 in detecting change because as above, the task appears to call upon knowledge in a larger number of aspects when compared to Lex30 even though this was not the original design of the task (Fitzpatrick & Clenton, 2017).

As for the third task, the PVLT only accepts one possible answer in response to each of its 90 sentence elicitation gaps. I suggest that the PVLT likely and potentially elicits a smaller number of responses than Lex30 and G_Lex because the PVLT demands specific elicitation as dictated by the predetermined responses that require attention to context in terms of semantic appropriateness and grammatical accuracy (e.g., for the missing gap in "Plants receive water from the soil through their ro….," both A2 and B2 levels respond with "*roots,"* which is likely more demanding than to provide single-word responses to Lex30 or responses with apparent attention to semantic appropriateness and grammatical accuracy as in

G_Lex. Accordingly, it is expected that the PVLT is the least sensitive of the three tasks to detect change.

Figure 7.1

Samples of Participants' Responses to Lex30, G_Lex, and the PVLT for A2 level

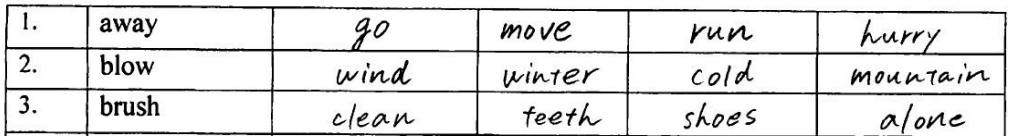
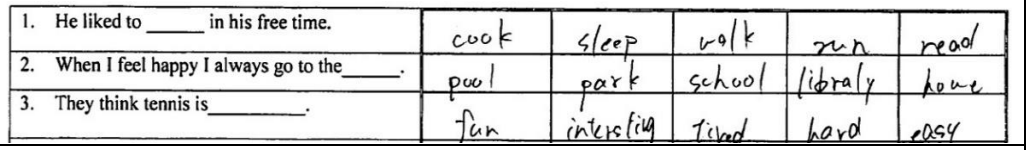
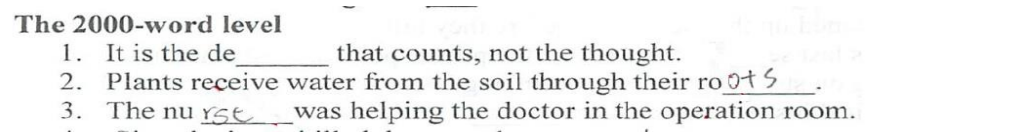
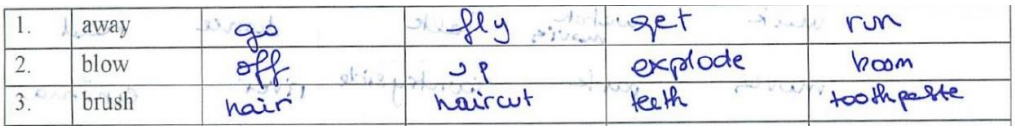
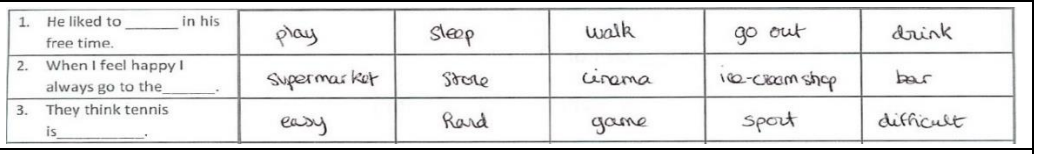
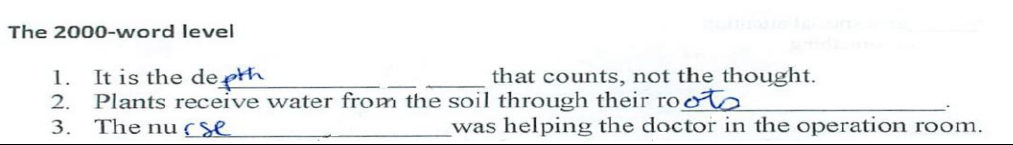
Task	A2 level (Chapter 4)
Lex30	
G_Lex	
PVLT	

Figure 7.2

Samples of Participants' Responses to Lex30, G_Lex, and the PVLT for B2 level

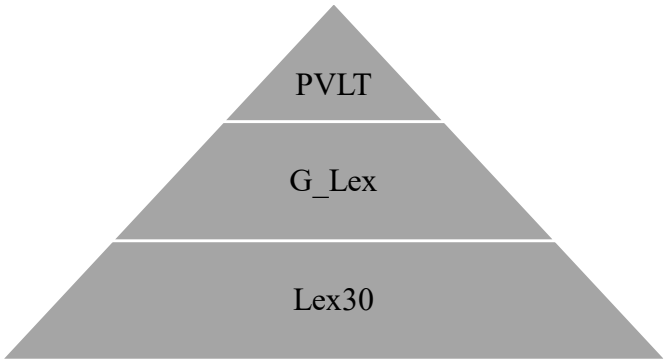
Task	B2 level (Chapter 5)
Lex30	
G_Lex	
PVLT	

These earlier hypotheses on task elicitation characteristics and the extent to which "context" might relate to task production (Lex30 > G_Lex > PVLT) form a theoretical implicational scale, whereby it is expected that Lex30 be more sensitive in detecting change

than both G_Lex and the PVLT; G_Lex is expected to be more sensitive in detecting change than the PVLT (Figure 7.3).

Figure 7.3

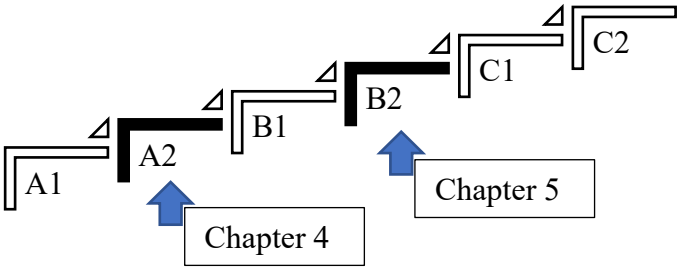
Suggested Implicational Scale of the Three Productive Vocabulary Tasks



The results from Chapters 4 and 5 allow me to justify the hypothesized implicational scale. Both chapters administrated two different versions of the same three vocabulary tasks (Lex30, G_Lex, and the PVLT) within the same three-month period. A main difference between Chapters 3 and 4 is the participant proficiency level (i.e., the ability to produce more infrequent words in context and contextually demanding tasks such as the PVLT and the LFP). The participants in Chapter 4 were CEFR A2 level, whereas in Chapter 5 they were at CEFR B2 (Figure 7.4). The Common European Framework of Reference (CEFR; Council of Europe, 2001), refers to A2 users as being able to produce a "simple terms" description for "aspects of his/her background, immediate environment and matters in areas of immediate need" (p. 24) while the B2 proficient users are able to "produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and independent disadvantages of various options" (p. 24).

Figure 7.4

Chapters 4 and 5 Participants' CEFR Levels of Proficiency



The findings from Chapter 4 (Table 7.1) are consistent with the hypothesized implicational relationship among the three tasks under investigation and according to which I expected Lex30 to be the most sensitive to change in productive vocabulary knowledge followed by G_Lex and then the PVLT (Figure 7.5). The findings from Chapter 4 also suggest that the change in productive lexical knowledge over the course of three months did not relate to the knowledge of semantic appropriateness and grammatical accuracy, which I hypothesized to be measured by the PVLT.

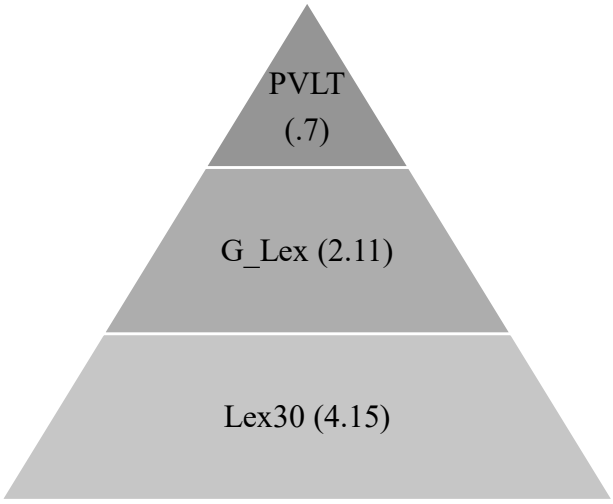
Table 7.1

Change across Time as Measured by Lex30, G_Lex, and the PVLT for A2 Level Participants

Task	Chapter 4 (A2 level)
Lex30	Significant change (4.15)
G_Lex	Significant change (2.11)
PVLT	No significant change (.7)

Figure 7.5

Implicational Scale for A2 Participants



Whereas the hypothesized implicational relationship (Lex30 > G_Lex > PVLT) appears to be consistent with the results found for CEFR A2 participants, the same cannot be said for the CEFR B2 level (G_Lex > Lex30 > PVLT) (Figure 7.6). The findings from Chapter 5 (Table 7.2) show significant changes over time for the G_Lex task but not for the PVLT or the Lex30 tasks among a group of B2 level participants. The findings from Chapter 5 also suggest that the change in productive lexical knowledge over the course of three months did not relate to the knowledge of form, which I hypothesized to be measured by the Lex30, or the knowledge of semantic appropriateness and grammatical accuracy, which I hypothesized to be measured by the PVLT.

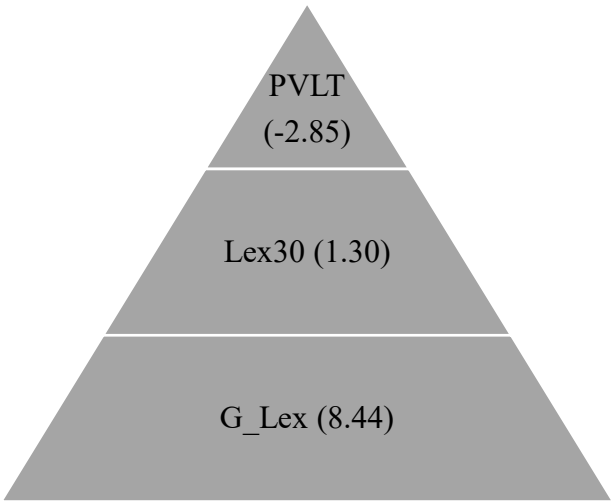
Table 7.2

Change across Time as Measured by Lex30, G_Lex, and the PVLT for B2 Level Participants

Task	Chapter 5 (B2 level)
Lex30	No significant change (1.30)
G_Lex	Significant change (8.44)
PVLT	No significant change (-2.85)

Figure 7.6

Implicational scale for B2 participants



7.1.2 Productive Vocabulary Tasks' Sensitivity to Detect Change According to a Proposed Theory of L2 Lexical Development

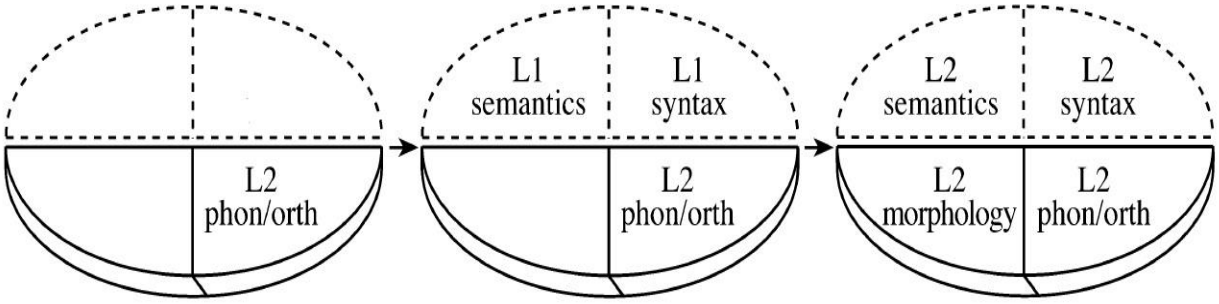
The previous section (7.2.1) discussed the extent to which the three productive vocabulary tasks (Lex30, G_Lex, and the PVLT) might detect change for different proficiency groups (A2 and B2). The proposed implicational scale suggests that Lex30 might be more sensitive (elicits more scoring responses) in detecting productive vocabulary knowledge change than both G_Lex and the PVLT, and that the G_Lex task might be more sensitive to detecting productive vocabulary knowledge change than the PVLT. The

implicational scale is argued to be accurate for those participants at the A2 level but not for those participants at the B2 Level. For the B2 level, G_Lex has shown to be more sensitive (i.e., elicits a greater proportion of scoring responses) than both Lex30 and the PVLT. Such lack of consistency in the hypothesized implicational scale with the B2 level suggests that factors related to lexical development (i.e., the ability to use specific words in context) might impact the extent to which tasks are sensitive to change.

To address the previous implicational scale inconsistency of G_Lex being more sensitive in detecting change than Lex30 with the B2 level, I turn to theory (Jiang, 2000; 2002; 2004) for support. Jiang's theoretical model of L2 lexical development introduces "three stages of L2 lexical development" as shown in Figure 7.7. The three stages describe how a specific word might develop in the learning process rather than how "lexical competence" develops as a whole. The first stage is "the formal stage when a lexical entry with formal specifications is established" (p. 47). At this initial stage, little content is established, mainly the word form and its L1 translation equivalent. The second stage is "the first language lemma mediation stage when the lemma information of the L1 counterpart is copied into the L2 lexical entry and mediates L2 word use" (p. 47). At the second stage, semantic and syntactic information of an L2 word might be copied from the L1 equivalent to form lexical entries of L2 forms and L1 lemmas. The third stage is "the L2 integration stage when semantic, syntactic, morphological specifications are integrated into the lexical entry" (p. 47). The third stage represents a full lexical development. An L2 lexical entry will be quite similar in "semantic, syntactic, and morphological specifications" to an L1 lexical entry.

Figure 7.7

Jiang's (2000) "Three Stages of Word Development"

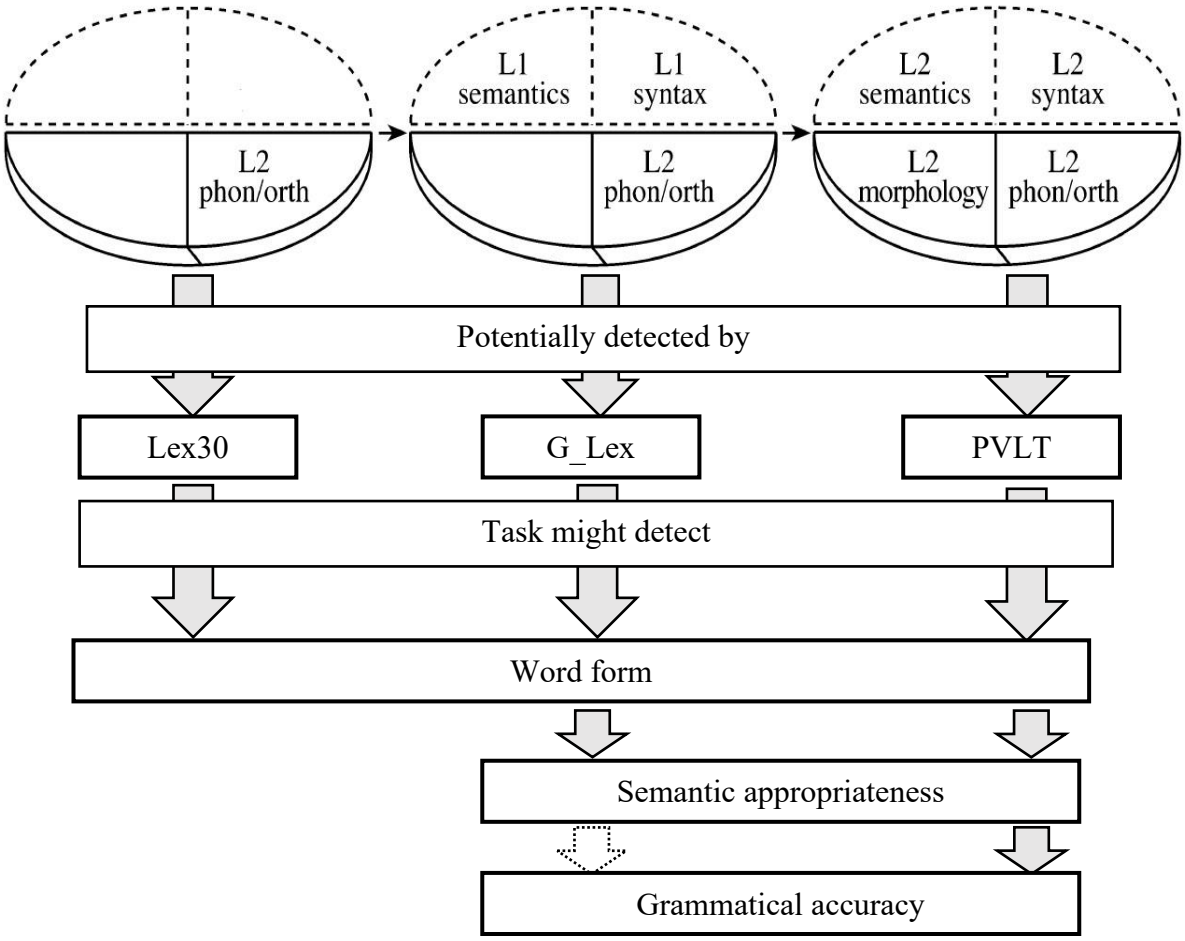


Note: Reprinted from "Lexical Representation and Development in a Second Language," by Jiang, N., 2000, *Applied Linguistics Journal*, 21(1), p. 54. Copyright 2000 by Oxford University Press.

I assume that such lexical development (words transit from one stage to another) might be manifest in the extent to which productive vocabulary tasks detect change for different proficiency levels. The three productive vocabulary tasks (Lex30, G_Lex, and the PVLT) and their characteristics (e.g., the extent to which tasks encourage task-takers to consider context) as introduced in Section 3.5, might tap into the different stages of Jiang's (2000) lexical development (Figure 7.8). Lex30 (2.1.3) might relate to word form knowledge, which might overlap with the first stage of Jiang's lexical development ("the formal stage"). G_Lex (2.1.4) might relate to word form, semantic appropriateness (and potentially grammatical accuracy) in context, which might overlap with the second stage of Jiang's lexical development ("the first language lemma mediation stage"). The PVLT (2.1.2) might relate to word form, semantic appropriateness, and grammatical accuracy in context, which might overlap with the third stage of Jiang's lexical development ("the L2 integration stage").

Figure 7.8

Productive Vocabulary Knowledge Tasks and Jiang's (2000; 2002; 2004) "Three Stages of Word Development"



* The dotted arrow from G_Lex indicates potential grammatical accuracy.

Note: The three stages of word development are adapted from "Lexical Representation and Development in a Second Language," by Jiang, 2000, *Applied Linguistics Journal*, 21(1), p. 54. Copyright 2000 by Oxford University Press.

These earlier hypothesized aspects of lexical development and the extent to which they might relate to task characteristics appear to indicate a theoretical "developmental scale" in which tasks might only detect development at specific lexical stages. Therefore, (i) if we are to see a change, then we might only see change within a test–retest per task and (ii) if we

run a test–retest for multiple different tasks (i.e., Lex30, G_Lex, and the PVLT), then the change might mainly occur within one specific task (stage of word development) as illustrated below for both A2 and B2 proficiency levels.

Although the A2 participants' lexicon might contain words at different stages of development, or even in transition from one stage to another, I assume that their L2 lexicon might be mainly at the first stage of word development ("the formal stage"). The formal stage was hypothesized to overlap with word form knowledge that might be tapped by Lex30. Lex30 allows task-takers to showcase the productive vocabulary knowledge "activated" by each of the 30 cues (Meara, 1997), thus facilitating the production of responses about which respondents have (more or less) partial knowledge. Because the responses to Lex30 are not presented within any context, the knowledge of items produced cannot be claimed to be anything other than at the level of threshold level knowledge (i.e., "the formal stage," Jiang's first stage of word development). Building on the two previous assumptions (i) that A2 participants' lexicon might be mainly at the first stage of Jiang's word development and (ii) responses to Lex30 might represent examples of Jiang's first stage of word development, I expect that a test–retest of Lex30 might show greater sensitivity to change than either G_Lex or the PVLT for the A2 participants.

By contrast, although the B2 participants' lexicon might contain words at different stages of development, or even in transition from one stage to another, I assume that their L2 lexicon might be mainly at the second stage of word development ("the first language lemma mediation stage"). The first language lemma mediation stage was hypothesized to overlap with word form, semantic appropriateness, and potential grammatical accuracy knowledge that might be tapped by G_Lex. Although G_Lex scoring is based solely on the frequency of items provided (and not on semantic appropriateness or grammatical accuracy), task responses indicated that they appear to match the semantic appropriateness (with potential

grammatical accuracy) of the sentence elicitation cues (i.e., "the first language lemma mediation stage," Jiang's second stage of word development). Building on the two earlier assumptions of (i) B2 participants' lexicon might be mainly at the second stage of Jiang's word development and (ii) responses to G_Lex might represent examples of Jiang's second stage of word development; I expect that a test–retest of G_Lex might show greater sensitivity to change than either Lex30 or the PVLT for the B2 participants.

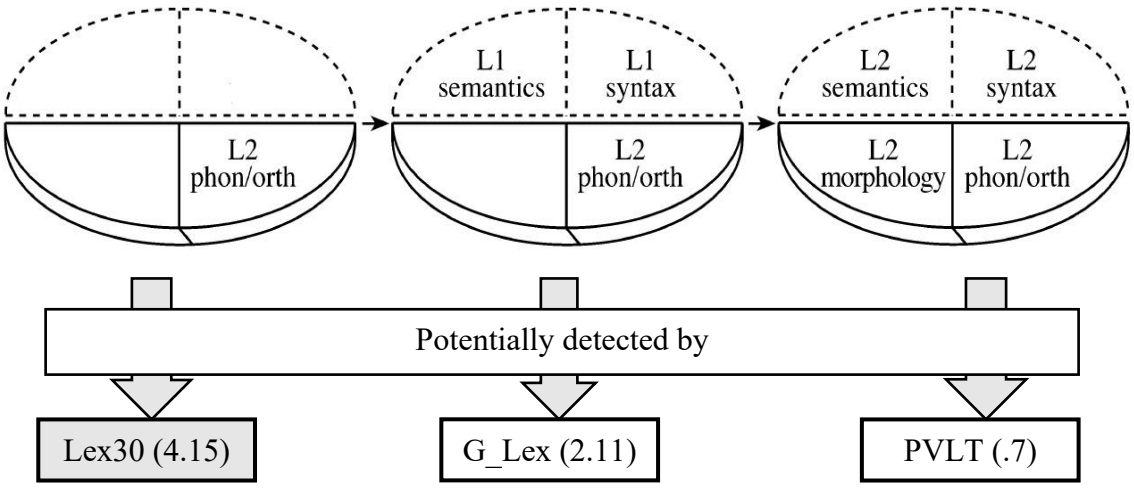
Furthermore, my discussion of task sensitivity to detect change according to Jiang's stages of word development has also assumed that the PVLT might show less sensitivity to change than either Lex30 or G_Lex for the two proficiency levels (A2 and B2) that were the focus of the experimental chapters (4 and 5). This assumption is made largely on the basis that (i) unlike the Lex30 and G_Lex tasks, the PVLT only accepts one response to each of its 90 sentence elicitation gaps, resulting in the range of activation events being more limited, and (ii) in order to provide a PVLT scoring response, participants need knowledge of semantic appropriateness as well as grammatical accuracy meaning that responses to the PVLT might represent examples of the third level of Jiang's word development ("the L2 integration stage"). A change in the third level of word development (i.e., more words transit from the second level to the third level of word development), as argued by Jiang (2000), might require a longer time (more than the three months of the current study). I, thus, expect that a test-retest of the PVLT might show less sensitivity in detecting change compared with Lex30 and G_Lex for the two proficiency levels (A2 and B2). I will return to this thread in the concluding chapter where I reflect on the extent to which the three productive vocabulary knowledge tasks might detect change for different proficiency levels.

To empirically test the hypothesized theoretical developmental scale of (Lex30 > G_Lex > PVLT) for the A2 learners, and (G_Lex > Lex30 > PVLT) for the B2 participants, I return to the findings from my fourth and fifth experimental chapters. The findings from

Chapter 4 show that the difference in the A2 participant scoring percentages between the two test times is greater for Lex30 than for the G_Lex and PVLT tasks. These findings are consistent with the hypothesized developmental scale among the three tasks under investigation and according to which I expected Lex30 to be more sensitive to detect change than the other two tasks. The findings from Chapter 4 also suggest that the change in productive lexical knowledge for a group of A2 level over the course of three months might advance in knowledge of word form, which is hypothesized to be the first stage of word development (Figure 7.9).

Figure 7.9

Hypothesized Task Developmental Scale for A2 Participants in Relation to Jiang's (2000; 2002; 2004) "Three Stages of Word Development"



* Number in parentheses is the average change in percentage (Test 2 − Test 1)

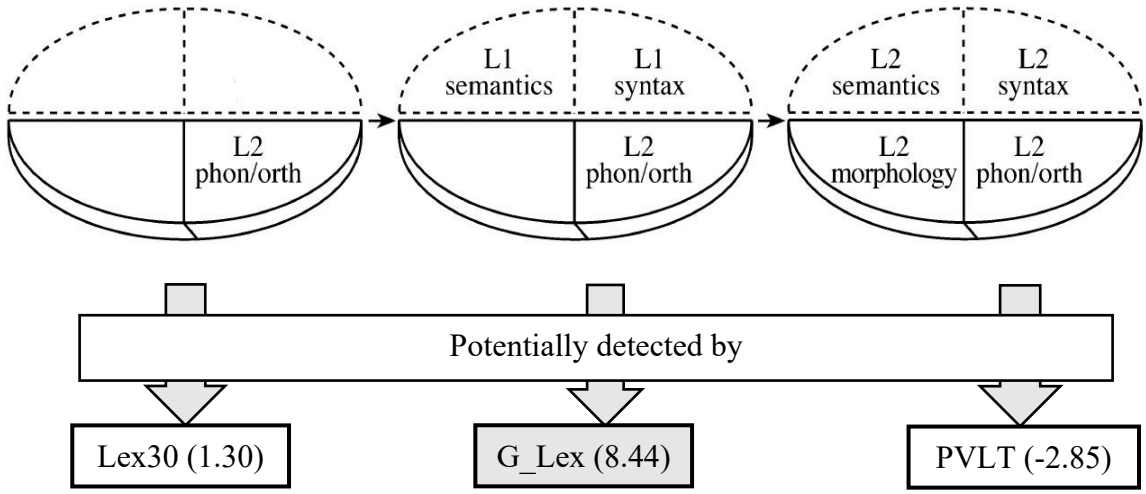
** Lex30 has shown the greatest increase in infrequent words for this group of A2 participants.

Note: The three stages of word development are adapted from "Lexical Representation and Development in a Second Language," by Jiang, N., 2000, *Applied Linguistics Journal*, 21(1), p. 54. Copyright 2000 by Oxford University Press.

The findings from Chapter 5 show that the difference in the B2 participant scoring percentages between the two test times is greater for G_Lex than for Lex30 and the PVLT. These findings are consistent with the hypothesized lexical stage task-dependent developmental scale among the three tasks under investigation and according to which I expected G_Lex to be more sensitive to detect change than the other two tasks. The findings from Chapter 5 also suggest that the change in productive lexical knowledge for a group of B2 level learners over the course of three months might advance in knowledge of semantic appropriateness (and potentially in grammatical accuracy), which is hypothesized to be the second stage of word development (Figure 7.10).

Figure 7.10

Hypothesized Task Developmental Scale for B2 Participants in Relation to Jiang's (2000; 2002; 2004) "Three Stages of Word Development"



* Number is parentheses is the average increase in percentage (Test 2 − Test 1)

** G_Lex has shown the greatest increase in infrequent words for this group of B2 participants.

Note: The three stages of word development are adapted from "Lexical Representation and Development in a Second Language," by Jiang, N., 2000, *Applied Linguistics Journal*, 21(1), p. 54. Copyright 2000 by Oxford University Press.

7.1.3 Productive Vocabulary Tasks' Sensitivity to Detect Consistent Change

The previous sections (7.1.1 and 7.1.2) investigated the sensitivity of productive vocabulary tasks in detecting change based on two different perspectives. The first perspective is task characteristics: an implicational scale (Lex30 > G_Lex > PVLT) based on task characteristics in Section 3.5 was created and hypothesized to account for A2 level participant responses but not for the B2 group. The second perspective is lexical development: a developmental scale based on the three stages of word development as proposed by Jiang (2000; 2002; 2004), which was created and argued to be accurate for A2 level participants (Lex30 > G_Lex > PLVT) and B2 level participants (G_Lex > Lex30 > PVLT).

While the change measured by each task might vary according to the task characteristics (hypothesized aspects of construct being measured (i.e., form, semantic appropriateness, or grammatical accuracy) and participants' proficiency (stage of lexical development), G_Lex has shown to be the only task (when compared with Lex30 and the PVLT) that detects consistent (significant) change between the two different proficiency groups in three short-term (three-month) longitudinal studies (A2 in Chapters 4 and 6, and B2 in Chapter 5, Table 7.3). Such consistency in detecting significant change might relate to the word knowledge changes expected in the short time (i.e., improvements in the ability to use a word form that is semantically appropriate as well as potentially in a grammatically accurate manner).

Table 7.3

Detecting Productive Vocabulary Knowledge Change across Time as Measured by Lex30, G_Lex, and the PVLT for Chapters 4–6

Task	Chapter 4 (A2 level)	Chapter 5 (B2 level)	Chapter 6 (A2 level)
Lex30	Significant change (4.15)	No significant change (1.30)	No significant change (1.36)
G_Lex*	Significant change (2.11)	Significant change (8.44)	Significant change (5.10)
PVLT	No significant change (.7)	No significant change (-2.85)	Significant, change (−3.93)

* Only G_Lex detected consistent change across the three longitudinal experiments for both proficiency levels (A2 and B2).

While there remains some degree of debate to propose a single productive vocabulary task that might be sensitive enough to detect change for different proficiency levels, G_Lex might provide this function for these two proficiency groups. This suggestion of G_Lex sensitivity to detect change might offer some useful insights to the discussion about how to assess various (e.g., pre- sessional) short-term (i.e., 12-week) language programs. Pre-sessional language programs are usually designed to develop language skills for students who do not speak the target language as their first language. These language courses are usually conducted before the main course at a host university and can last 6–41 weeks (e.g., in the USA, UK, Australia, NZ, and Canada).

Having established a theoretical background that relates to how the three productive vocabulary knowledge tasks might detect proficiency changes in productive vocabulary knowledge, I return to considering Chapters 5 and 6 in order to explore pedagogical implications for the three productive vocabulary tasks (Lex30, G_Lex, and PVLT) in relation to "global measures." The experiment reported in Chapter 5 showed that productive vocabulary knowledge tasks might not be consistent in detecting change using an IELTS writing task for a group of B2 participants. The experiment reported in Chapter 6 showed that productive vocabulary knowledge tasks might not detect change in aspects of IELTS speaking for a group of A2 participants.

7.2 Productive Vocabulary Tasks' Sensitivity to Detect Change in Frequent Words Use in IELTS Writing and Speaking

Re-examining the IELTS data (writing and speaking tasks) from both Chapters 5 and 6, this section (7.2) aims to link the significant changes in productive vocabulary tasks to global production proficiency measures (i.e., IELTS). Each year, IELTS is administered to over 3.5 million test-takers around the world (IELTS, 2019b), with a steadily increasing number of test-takers in Japan. IELTS speaking and writing (productive skills) tend to be more challenging for test-takers than IELTS listening and reading (receptive skills). As an example, L1 Japanese L2 English learners tend to receive lower average scores in IELTS speaking and writing (5.5 out of 9 for each) than IELTS listening and reading (5.9 and 6.1 out of 9, respectively) (IELTS, 2019c).

Accordingly, Section 7.2 explores the extent to which the three productive vocabulary tasks detect change in the use of infrequent words in response to the (i) IELTS Writing Task 2 and (ii) the three IELTS speaking tasks. I first discuss the extent to which the three productive vocabulary tasks detect the infrequent words produced in response to IELTS Writing Task 2 for B2 participants. Second, I discuss the extent to which the three productive

vocabulary tasks detect the infrequent words produced in response to IELTS speaking tasks (for A2 participants). Third, by building on the first two threads of this section, I then explore several potential implications that might relate to pedagogical implications for both the IELTS writing and speaking tasks.

7.2.1 Comparing the Vocabulary Produced in Response to the Productive Vocabulary Knowledge Tasks and the IELTS Writing Task 2 for B2 Participants

The current section aims to explore the extent to which the three productive vocabulary tasks might detect change in infrequent words use in response to the IELTS Writing Task 2. The IELTS Writing Task 2 requires test-takers to write a response to an argument or problem within 40 minutes. Topics for the IELTS Writing Task 2 vary but all require task-takers to present clear, detailed, well-defended points of view. The IELTS writing rubric for Task 2 evaluates four different aspects of writing band descriptors: task response, coherence and cohesion, lexical resource, and grammatical range and accuracy. Each writing aspect consists of nine bands with band 1 being the lowest score and band 9 the highest. The nine band descriptors of the lexical resource aspect are mainly based on the learner's ability to use a wide range of vocabulary with "less common lexical items" as illustrated by the lexical resource in the IELTS writing band descriptors (Figure 7.11).

Figure 7.11

Lexical resource in IELTS task writing band descriptors (public version)

(IELTS, 2021a)

Band 1	▪ can only use a few isolated words
Band 2	▪ uses an extremely limited range of vocabulary; essentially, no control of word formation and/or spelling
Band 3	▪ uses only a very limited range of words and expressions with very limited control of word formation and/or spelling ▪ errors may severely distort the message
Band 4	▪ uses only basic vocabulary, which may be used repetitively or be inappropriate for the task ▪ has limited control of word formation and/or spelling; errors may cause strain for the reader
Band 5	▪ uses a limited range of vocabulary, but this is minimally adequate for the task ▪ may make noticeable errors in spelling and/or word formation that may cause some difficulty for the reader
Band 6	▪ uses an adequate range of vocabulary for the task ▪ attempts to use less common vocabulary but with some inaccuracy ▪ makes some errors in spelling and/or word formation, but they do not impede communication

Band 7	▪ uses a sufficient range of vocabulary to allow some flexibility and precision ▪ uses less common lexical items with some awareness of style and collocation ▪ may produce occasional errors in word choice, spelling, and/or word formation
Band 8	▪ uses a wide range of vocabulary ▪ fluently and flexibly conveys precise meanings ▪ skillfully uses uncommon lexical items but there may be occasional inaccuracies in word choice and collocation ▪ produces rare errors in spelling and/or word formation
Band 9	▪ uses a wide range of vocabulary with very natural and sophisticated control of lexical features; rare minor errors occur only as "slips"

Note: Adapted from https://takeielts.britishcouncil.org/sites/default/files/ielts_task_2_writing_band_descriptors.pdf. Copyright 2021 by IELTS.

Evidence exists within the literature of attempts to elucidate relationships between large-scale testing suites and vocabulary knowledge (e.g., Johnson et al., 2016), as reviewed in Section 2.2.3. Johnson et al. (2016) investigated the relationships among three different vocabulary measures (receptive vocabulary size (Nation & Beglar, 2007); aural vocabulary (Fountain & Nation, 2000); and productive levels (Laufer & Nation, 1999)) and a TOEFL (ETS, 2005) independent writing task. The study reported significant correlations ($r = .38$, $p = .003$) between productive vocabulary test scores, "holistic writing quality scores," and "L2 writing performance." While Johnson et al. (2016) is useful because it supports an

investigation of the potential relationships among multiple vocabulary tasks and writing performance on a "high-stakes" writing test (i.e., TOEFL), there are significant differences. Here, the experimental data are based on (i) longitudinal data, (ii) multiple productive vocabulary task (G_Lex, Lex30, and PVLT) scores, and (iii) less common lexical items as tapped by the IELTS Writing Task 2.

Taking these differences into account, Chapter 5 investigated the extent to which different productive vocabulary tasks detect significant changes in relation to infrequent words produced in the IELTS Writing Task 2 for a group of B2 participants. Extending Chapter 5's discussion (5.4), the current section explores the extent to which the three productive vocabulary tasks (Lex30, G_Lex, and the PVLT) might detect the infrequent words change in response to the LFP (IELTS Writing Task 2).

Consideration of Chapter 5 data with Pearson correlation analyses among the three productive vocabulary tasks scores and a count of the infrequent (i.e., non-1k) words produced in response to the IELTS Writing Task 2 for both testing times (Tables 7.4 and 7.5) reveals two important findings. First, the three productive vocabulary task (Lex30, G_Lex, and the PVLT) scores correlate significantly ($p < .001$) with the count of the infrequent words produced in response to the IELTS Writing Task 2 for this group of B2 learners. Second, although the three productive vocabulary task scores significantly correlate with the number of infrequent words produced in response to the IELTS Writing Task 2 in the first testing time, only the PVLT task scores (compared with Lex30 and G_Lex) correlate significantly with the number of infrequent words produced in response to the IELTS Writing Task 2 at both testing times (Time 1: $r = 0.450$, $p < .01$; Time 2: $r = 0.406$, $p < .01$). Such a moderate consistent relationship between the PVLT and IELTS Writing Task 2 might relate to task characteristics.

As illustrated in Sections 3.5 and 7.1, Lex30 examines the ability to produce individual items activated in response to single-word stimuli. It is hypothesized that this elicits a minimal quality of knowledge (i.e., word form). Unlike Lex30, G_Lex encourages test-takers to consider context when providing their responses as it is hypothesized to stimulate responses with potential semantic appropriateness and grammatical accuracy. Unlike Lex30 and G_Lex, the PVLT only accepts one response to each of its 90-sentence letter-prompt gaps, which is hypothesized to elicit responses with both semantic appropriateness and grammatical accuracy.

Perhaps, like the PVLT, participants' responses to the IELTS Writing Task 2 might represent knowledge of semantic appropriateness and grammatical accuracy. In other words, words are likely to be produced in the IELTS Writing Task 2 only if the task-taker feels he or she has, at least to some extent, semantic and grammatical mastery of their use. Laufer and Nation (1995) noted that words were generally used in a correct way in essay responses to a discussion question (p. 315).

The hypothesized relationship between the elicited vocabulary by both the PVLT and the IELTS Writing Task 2 (as illustrated in Tables 7.4 and 7.5) might provide evidence of classroom use. It is suggested that a change in the PVLT scores might moderately reflect a change in the infrequent words produced in response to the IELTS Writing Task 2 for a group of B2 participants.

Table 7.4

Correlations between the Three Productive Vocabulary Task Mean Scores and Number of Infrequent Words Produced in Response to the IELTS Writing Task 2 for Test 1 for a Group of B2 Participants

Task	Lex30	G_Lex	PVLT
G_Lex	0.645***		
PVLT	0.612***	0.526 ***	
IELTS Writing Task 2	0.489***	0.316*	0.450**

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

Table 7.5

Correlations between the Three Productive Vocabulary Task Mean Scores and Number of Infrequent Words Produced in Response to the IELTS Writing Task 2 for Test 2 for a Group of B2 Participants

Task	Lex30	G_Lex	PVLT
G_Lex	0.548***		
PVLT	0.586***	0.617***	
IELTS Writing Task 2	0.110	0.018	0.406**

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

To obtain a picture of both productive language skills, an investigation of the extent to which the three productive vocabulary tasks might detect infrequent words produced in response to IELTS speaking tasks might be needed. The next section, therefore, explores potential relationships between the infrequent words produced in response to IELTS speaking tasks and the infrequent words elicited by the same three productive vocabulary tasks (Lex30, G_Lex, and the PVLT).

7.2.2 Comparing the Infrequent Words Produced in Response to the Productive Vocabulary Knowledge Tasks and the IELTS Speaking Tasks 2 for A2 Participants

The current section aims to explore the extent to which the three productive vocabulary tasks might detect change to infrequent word use in response to the three tasks of the IELTS speaking test. The IELTS speaking test consists of three parts and lasts for approximately 13–14 minutes. Part 1 requires test-takers to respond to informal interview questions covering family, work, study, hobbies, etc., and lasts for approximately five minutes. Part 2 requires test-takers to explain and describe a specific topic presented on a card or prompt provided by their examiner. Test-takers are given one minute to prepare responses and approximately two minutes to respond. Part 3 requires test-takers to respond to a variety of questions such as opinion giving, comparing, predicting the future, talking about the past, talking about people in general, etc., and lasts for approximately five minutes.

The IELTS speaking rubric evaluates four different aspects of speaking: fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation. Each aspect consists of nine bands where band 1 represents the lowest score and band 9 the highest.. The nine bands of the lexical resource aspect in the IELTS speaking test are based on the ability to use a wide range of vocabulary with “less common” words as described in Figure 7.12.

Figure 7.12

Lexical Resource in the IELTS Speaking Band Descriptors (public version) (IELTS, 2021b)

Band 1	▪ no communication possible •no rateable language
Band 2	•only produces isolated words or memorised utterances
Band 3	•uses simple vocabulary to convey personal information •has insufficient vocabulary for less familiar topics
Band 4	•is able to talk about familiar topics but can only convey basic meaning on unfamiliar topics and makes frequent errors in word choice •rarely attempts paraphrase
Band 5	•manages to talk about familiar and unfamiliar topics but uses vocabulary with limited flexibility •attempts to use paraphrase but with mixed success
Band 6	•has a wide enough vocabulary to discuss topics at length and make meaning clear in spite of inappropriacies •generally paraphrases successfully
Band 7	•uses vocabulary resource flexibly to discuss a variety of topics •uses some less common and idiomatic vocabulary and shows some awareness of style and collocation, with some inappropriate choices •uses paraphrase effectively
Band 8	•uses a wide vocabulary resource readily and flexibly to convey precise meaning •uses less common and idiomatic vocabulary skilfully, with occasional inaccuracies •uses paraphrase effectively as required

Band 9	•uses vocabulary with full flexibility and precision in all topics •uses idiomatic language naturally and accurately

Note: Adapted from https://www.ielts.org/-/media/pdfs/speaking-band-descriptors.ashx?la=en. Copyright 2021 by IELTS.

Evidence exists within the literature of attempts to compare the vocabulary used in response to both vocabulary and fluency elicitation tasks (e.g., Clenton et al., 2021). Clenton et al. (2021), as reviewed in Section 2.2.2, investigated the relationships between productive vocabulary knowledge tapped by a productive vocabulary task (Lex30) and vocabulary used in experimental speaking fluency tasks. The study found significant correlations ($p < 0.05$) between the vocabulary used in both speaking and vocabulary tasks at word frequency levels 0 and 4 of the Academic Spoken Word List (ASWL). The following discussion, building on Clenton et al. (2021), presents significant differences based on (i) longitudinal data, (ii) multiple productive vocabulary task (G_Lex, Lex30, and PVLT) scores, and (iii) the infrequent words produced in response to a speaking test (i.e., IELTS).

Taking the stated differences into account, Chapter 6 investigated the extent to which different productive vocabulary tasks detect consistent changes in relation to measures of speaking fluency for the IELTS speaking tasks for a group of A2 participants. Extending Chapter 6's discussion (6.4), the current section explores the extent to which the three productive vocabulary tasks (Lex30, G_Lex, and the PVLT) might detect change to the use of infrequent words in response to the IELTS speaking test.

A re-exploration of Chapter 6 data with correlations between infrequent vocabulary produced in response to the productive vocabulary knowledge tasks and the IELTS speaking tasks (Table 7.6) shows three important findings. First, only G_Lex scores correlate significantly ($p < .01$) with the number of infrequent words produced in response to the

IELTS speaking Part 3 in the first testing time. Second, there is no significant correlation between the three productive vocabulary task scores and the number of infrequent words produced in response to the three parts of the IELTS speaking test in the second testing time. Third, there is no significant relationship between the three productive vocabulary task scores and the number of infrequent words produced in response to the combined three parts of the IELTS speaking test in both testing times.

Table 7.6

Correlations between the Three Productive Vocabulary Task Scores and Number of Infrequent Words Tapped by the IELTS Speaking Tasks in the Two Testing Times for A2 Participants

Tasks	Times	IELTS Speaking Part 1	IELTS Speaking Part 2	IELTS Speaking Part 3	IELTS Task Combined
Lex30	Time	0.098	−0.140	0.233	0.022
G_Lex	1	−0.018	−0.051	0.440**	0.138
PVLT		−0.061	−0.276	−0.025	−0.230
Lex30	Time	−0.057	0.224	−0.101	0.058
G_Lex	2	−0.100	0.022	−0.077	−0.078
PVLT		0.060	−0.005	−0.154	−0.048

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

Such an inconsistent relationship among the three productive vocabulary task scores and the number of infrequent words produced in response to different IELTS speaking tasks needs explaining. One potential explanation might relate to time, “in particular because of the strict time constraints under which the speaker needs to operate” (De Jong, 2016, p. 203). A speaker needs to rapidly and consistently translate his thoughts into “intelligible sounds” in

a short period of time (De Jong, 2016). Such a rapid and consistent process of translating thoughts into sounds might minimize the use of infrequent words, especially for lower-level proficiency participants. According to Jiang (2000), to rapidly produce words with semantic appropriateness and grammatical accuracy, words should reach the third level of word development ("the L2 integration stage"). For beginner level participants, I expect that most of their level 3 words might be within the 1k and, therefore, their oral vocabulary production might contain a greater proportion of highly frequent words.

In addition, and as illustrated in Table 7.7, the change (Time 2 - Time 1) of infrequent words use in IELTS speaking Part 2 was significantly higher than for Parts 1 and 3. Unlike Parts 1 and 3, Part 2 relaxes the time constraint by giving test-takers one minute to prepare followed by two minutes to speak without interruption from the examiner. Therefore, in general terms for this group of A2 participants, I assume that the less the time constraint in speaking tasks, the more infrequent words we might see in use. However, the previous discussion was for a single group of A2 participants. Subsequent studies of different proficiency levels (e.g., B2 and C2) might be needed to support the view the time allowed to test-takers might relate to infrequent words production in speaking tasks.

Table 7.7

Percentage of Infrequent Words (non-1k) in the Three Tasks of the IELTS Speaking Test for a Group of A2 Participants

Tasks	Time 1	Time 2	Time 2 - Time 1
IELTS Part 1	20.475 (8.351)	21.725 (7.400)	1.25
IELTS Part 2	19.625 (8.308)	25.200 (9.993)	5.575*
IELTS Part 3	20.075 (6.149)	22.075 (8.914)	2

Note: * $p < .05$, ** $p < .01$, *** $p < .001$ (Bonferroni correction). Numbers in parentheses are standard deviations.

To conclude, this section discussed (i) the inconsistent relationships among the three productive vocabulary task scores and the number of infrequent words produced in response to the IELTS speaking test, and (ii) how such inconsistency might relate to time constraint might yield various implications for IELTS test-takers.

7.2.3 Potential Implications for IELTS Writing and Speaking Courses

The previous sections (7.2.1 and 7.2.2) explored several potential relationships between the knowledge elicited by the three productive vocabulary tasks (Lex30, G_Lex, and the PVLT) and both the IELTS Writing Task 2 for B2 participants and IELTS speaking test for A2 participants. This section (7.2.3), builds on sections 7.2.1 and 7.2.2, to explore potential pedagogical implications for IELTS writing and speaking.

Section 7.2.1 discussed potential consistent relationships between the PVLT task scores and the infrequent words tapped by the IELTS Writing Task 2. Although moderate significant relationships among the three productive vocabulary tasks scores and the less common words in IELTS Writing Task 2 were found at the first testing time, only the PVLT has shown moderate significant relationship in both testing times (Time 1: $r = 0.450$, $p < .01$; Time 2: $r = 0.406$, $p < .01$). No significant relationship between either Lex30 or G_Lex tasks and the infrequent words in the IELTS Writing Task 2 was found in the second testing time.

Such a consistent moderate relationship between the PVLT and IELTS Writing Task 2 might provide pedagogical implications for the IELTS writing course. The IELTS writing course might take advantage of the freely available versions of the PVLT task to predict changes in participants' infrequent word use on IELTS Writing Task 2. As detailed in Section 7.2.1, the use of "uncommon lexical items" is a main descriptor for the lexical resource in IELTS Writing Task 2. In other words, the more infrequent words with semantic appropriateness and grammatical accuracy the task-takers may use in IELTS Writing Task 2, the better score they may receive.

The PVLT as a potential predictor of infrequent word use change on IELTS Writing Task 2, on the other hand, might save the amount of time needed to administer, transcribe, and process the writing task. Each PVLT task requires an average of 25 minutes to complete and a few minutes to process. There are three available, free to use versions of the PVLT task (provided in Appendix 2). The three PVLT tasks are also available online at: https://www.lextutor.ca/tests/levels/productive/.

However, this discussion was for a single group of B2 participants. Subsequent studies of different proficiency levels (e.g., C2) might be needed to support the use of the PVLT as a potential consistent predictor for the IELTS Writing Task 2 lexical resource score change.

Section 7.2.2 explored several inconsistent relationships between the three productive vocabulary task scores and the infrequent words tapped by the three IELTS speaking tasks. As illustrated in Table 7.7, only G_Lex has shown a moderate significant relationship (r = 0.440, $p < .01$) with the use of infrequent words in the IELTS speaking Part 3 in the first testing time. No significant relationship among the three productive vocabulary tasks and infrequent word use in the three parts of the IELTS speaking test was found in the second testing time.

Such a moderate significant relationship between G_Lex and IELTS speaking Part 3 might provide pedagogical implications for the IELTS speaking course. The IELTS speaking course might take advantage of the freely available versions of the G_Lex task in order to predict infrequent words use on the IELTS speaking test Part 3. As detailed in Section 7.2.2, the use of "uncommon lexical items" is a main descriptor for lexical resource in the IELTS speaking test. In other words, the more infrequent words with potential semantic appropriateness and grammatical accuracy the test-takers may use in IELTS speaking test, the better score they may receive.

Using the G_Lex to predict infrequent words on the IELTS speaking test Part 3, on the other hand, might save the amount of time needed to administer, transcribe, and process the speaking task. There are three available, free to use versions of the G_Lex task (provided in Appendix 4). However, the previous discussion was for a single group of A2 participants. Subsequent studies of different proficiency levels (e.g., B2 and C2) might be needed to support the use of the G_Lex as a potential predictor for the IELTS speaking test Part 3 lexical resource score.

7.3 Conclusion

Chapter 7 has drawn together the different strands of the thesis and discussed these in two broad sections. The first section addressed three issues related to productive vocabulary sensitivity to detect change. The first of these related to the extent to which productive vocabulary tasks' sensitivity to detect change might relate to task characteristics, suggesting that task sensitivity to change might be both task- and proficiency-based. I then discussed the extent to which productive vocabulary tasks sensitivity to detect change might relate to a proposed theory of L2 word development by Jiang (2000; 2002; 2004), suggesting that task sensitivity to change might relate to Jiang's three stages of word development. The third part of this first section examined productive vocabulary tasks' sensitivity to detect consistent change, suggesting that G_Lex might be a consistent productive vocabulary knowledge task for two proficiency levels (A2 and B2).

The second section (7.2) addressed three issues that relate to the extent to which productive vocabulary knowledge tasks might detect changes in infrequent words use in IELTS writing and speaking. The first issue discussed is how productive vocabulary tasks might detect changes in infrequent words use in IELTS Writing Task 2 for a group of B2 participants, suggesting that PVLT might be a consistent predictor for infrequent words use in IELTS Writing Task 2. The second issue discussed how productive vocabulary tasks might

detect change in infrequent words use in the IELTS speaking test for a group of A2 participants, suggesting that G_Lex might be a predictor for infrequent words in Part 3 of the IELTS speaking test. The third issue related to potential pedagogical implications for IELTS writing and speaking courses. This conclusion draws together the two main sections of Chapter 7 (7.1 and 7.2), which show that the extent to which productive vocabulary tasks detect change might be based on task, proficiency, word development, or even global measures (i.e., IELTS).

Chapter 8 Conclusion

The main purpose of my research relates to the extent to which productive vocabulary knowledge measures detect changes for two different proficiency groups. I began the thesis by presenting a review of four different vocabulary tasks (LFP, the PVLT, Lex30, and G_Lex). I then reviewed papers on how the knowledge elicited by vocabulary tasks might tap aspects that relate to speaking and writing skills. The last section in Chapter 2 reviewed several longitudinal papers on aspects of productive vocabulary development. The reviews in Chapter 2 suggested that further longitudinal productive vocabulary studies might be needed to (i) investigate productive vocabulary knowledge change over time and (ii) explore relationships between productive vocabulary knowledge and different aspects of writing and speaking skills. Thus, in my experimental chapters (3-6), I attempted to respond to those calls (Table 8.1).

In Chapter 3, I investigated the extent to which productive vocabulary knowledge varies according to the task. The cross-sectional study (n = 105 A2 participants) compared a widely cited measure (the Productive Vocabulary Levels Test (PVLT)) with two tasks from Fitzpatrick and Clenton (2017): Lex30 and G_Lex. The study reported that productive vocabulary knowledge is task-dependent for a group of A2 proficiency participants.

In Chapter 4, I investigated the extent to which results from a longitudinal study suggest that productive vocabulary knowledge tasks might detect change for a group of A2 proficiency participants. The longitudinal study (n = 100 A2 participants) measured participants' vocabulary knowledge with the same three productive vocabulary knowledge tasks from Chapter 3 (PVLT, Lex30, and G_Lex) at two times (0 and 3 months). The study reported that productive vocabulary change varies according to task for a group of A2 proficiency participants.

In Chapter 5, I investigated (i) the extent to which results from a longitudinal study suggest that productive vocabulary knowledge tasks might detect change for a group of B2 proficiency participants (a higher proficiency level than the A2 level reported in Chapter 4) and (ii) the extent to which results from a longitudinal study suggest that productive vocabulary knowledge change (as tapped by different productive vocabulary tasks) might be consistent with the use of less frequent words in a writing task. The longitudinal study (n = 50 B2 participants) measured participants' vocabulary knowledge with the same three productive vocabulary knowledge tasks from Chapters 3 and 4 (PVLT, Lex30, and G_Lex), but with an additional productive vocabulary writing task (the LFP), at two times (0 and 3 months). The study reported that (i) productive vocabulary change varies according to the task for a group of B2 proficiency participants and (ii) productive vocabulary knowledge change, as detected by different productive vocabulary tasks, inconsistently relate to infrequent words in writing for the same group of B2 proficiency participants.

In Chapter 6, I investigated the extent to which results from a longitudinal study suggest that productive vocabulary knowledge tasks scores are consistently related to measures of second language oral fluency. The longitudinal study (n = 45 A2 participants) explored potential relationships between the vocabulary knowledge detected by the same three vocabulary tasks from Chapters 3-5 (PVLT, Lex30, and G_Lex) and aspects of speaking fluency at two times (0 and 3 months). The study reported that aspects of L2 oral fluency inconsistently relate to productive vocabulary knowledge task scores for a group of A2 proficiency participants.

Table 8.1

Experimental Chapters (3–6)

Chapter	Type	*N*	Proficiency	Goals	Findings
3	Cross-sectional	105	CEFR A2	Investigate the extent to which productive vocabulary knowledge might vary according to task for a group of A2 proficiency participants.	Productive vocabulary knowledge is task-dependent.
4	Longitudinal*	100	CEFR A2	Investigate the extent to which productive vocabulary knowledge tasks might detect change for a group of A2 proficiency participants.	Productive vocabulary knowledge change varies according to the task.
5	Longitudinal*	50	CEFR B2	(i) Investigate the extent to which productive vocabulary knowledge tasks might detect change for a group of B2 proficiency participants.	(i) Productive vocabulary knowledge change varies according to the task and L2 proficiency.

				(ii) Investigate the extent to which productive vocabulary knowledge tasks score changes might consistently relate to aspects of writing for the same group of B2 proficiency participants.	(ii) productive vocabulary knowledge changes, as tapped by different productive vocabulary tasks, inconsistently relate to the production of infrequent words in writing
6	Longitudinal*	45	CEFR A2	Investigate the extent to which productive vocabulary knowledge tasks scores are consistently related to second language oral fluency for a group of A2 proficiency participants.	Aspects of L2 oral fluency inconsistently relate to productive vocabulary knowledge task scores.

* “Most recent longitudinal SLA studies span anywhere between three or four months and up to six years” (Ortega & Iberri-Shea, 2005, p. 37).

8.1 Study Claims

Following the experimental chapters (3-6), Chapter 7 attempted to draw together the different strands of the thesis to discuss six broad issues as summarized in three main claims. First, the extent to which productive vocabulary tasks can detect productive vocabulary knowledge change might be both implicational and developmental. Second, the extent to which productive vocabulary tasks can detect consistent productive vocabulary knowledge change for two different proficiency groups might be task-dependent. Third, the extent to which productive vocabulary tasks can detect productive vocabulary knowledge change on IELTS writing and speaking tasks might be task-dependent.

8.1.1 Detecting Changes in Productive Vocabulary Knowledge

The extent to which the three different productive vocabulary tasks (Lex30, G_Lex, and the PVLT) can detect change for two different proficiency groups might be both implicational- and developmental-based. The implicational-based change might relate to the quantitative change of specific aspect of productive vocabulary knowledge (e.g., form, semantic appropriateness, or grammatical accuracy) as detected by tasks with different elicitation characteristics. Tasks are hypothesized to tap into different aspects of the productive vocabulary construct: Lex30 is hypothesized to measure word form; G_Lex is hypothesized to measure the ability to use word form with semantic appropriateness and potential grammatical accuracy in context; and the PVLT is hypothesized to measure the ability to use word form with both semantic appropriateness and grammatical accuracy in context. In other words: (i) if Lex30 detects change, then such change might reflect a quantitative change in "'word form" as hypothesized to be tapped by Lex30; (ii) if G_Lex detects change, then such change might reflect a quantitative change in "semantic appropriateness with potential grammatical accuracy" as hypothesized to be tapped by G_Lex; and (iii) if the PVLT detects change, then such change might reflect a quantitative

change in “semantic appropriateness and grammatical accuracy” as hypothesized to be tapped by the PVLT.

The developmental-based change, on the other hand, might relate to the qualitative change of productive vocabulary knowledge (i.e., the ability to produce words in context and to perform contextually demanding tasks) as illustrated by Jiang’s (2000) word development theory: tasks are hypothesized to tap into different levels of Jiang’s word development stages; Lex30 is hypothesized to tap into Jiang’s first level of word development (“the formal stage”); G_Lex is hypothesized to tap into Jiang’s second level of word development (“the first language lemma mediation stage”); and the PVLT is hypothesized to tap into Jiang’s third level of word development (“the L2 integration stage”). In other words: (i) if Lex30 detects change, then such change might reflect a qualitative change in “the formal stage” as hypothesized to be tapped by Lex30; (ii) if G_Lex detects change, then such change might reflect a qualitative change in “the first language lemma mediation stage” as hypothesized to be tapped by G_Lex; and (iii) if the PVLT detects change, then such change might reflect a qualitative change in “the L2 integration stage” as hypothesized to be tapped by the PVLT.

8.1.2 Detecting Consistent Changes in Productive Vocabulary Knowledge

The extent to which the three different productive vocabulary tasks (Lex30, G_Lex, and the PVLT) can detect consistent change over a short time (i.e., three months) for two different proficiency groups (A2 and B2) might be task-dependent. G_Lex was the only task (when compared with Lex30 and the PVLT) that was able to detect consistent change between the two proficiency groups (A2 in Chapter 4 and B2 in Chapter 5) in two short-term (three-month) longitudinal studies. Lex30 detected consistent change with A2 level (Chapter 4) but inconsistent change with the B2 level (Chapter 5). The PVLT detected inconsistent changes with both A2 and B2 levels (Chapters 4 and 5).

8.1.3 Detecting Productive Vocabulary Knowledge Change in IELTS Writing and Speaking Tasks

The extent to which individual tasks can detect productive vocabulary knowledge change in IELTS writing and speaking tasks might be task-dependent. The PVLT was the only task (when compared with Lex30 and G_Lex) that was able to detect the infrequent words use change in the IELTS Writing Task 2 for a group of B2 participants. Both Lex30 and G_Lex scores were significant for infrequent words used in IELTS Writing Task 2 in the first testing time, but insignificant in the second testing time.

G_Lex was the only task (when compared with Lex30 and the PVLT) that was able to detect infrequent words used in the IELTS speaking test Part 3 for a group of A2 participants. G_Lex score was significant with the infrequent words in the IELTS speaking test Part 3 in the first testing time, but insignificant in the second testing time. Both Lex30 and the PVLT scores were insignificant with the three IELTS speaking tasks in both testing times.

8.2 Limitations and Future Research

There remain issues that might need to be addressed and are related to vocabulary knowledge tasks and their ability to detect changes. Accordingly, I have identified three main objectives for future research:

(i) To explore the extent to which multiple productive vocabulary knowledge tasks detect change for higher proficiency levels. The current studies were for groups of A2 and B2 proficiency levels. I assume that productive vocabulary knowledge tasks might be inconsistent in detecting change for advanced proficiency levels (higher than the proficiency levels reported in my studies), e.g., C1 and C2. Jiang (2000) stated that the transition from the second word level (L1 lemma mediation) to the third (highest) word level (L2 integration) may take a long time and might not even happen for most words (p. 55).

(ii) To explore the extent to which individual productive vocabulary knowledge tasks detect different knowledge as indicated by the IELTS speaking and writing descriptors (e.g., "grammatical range and accuracy"). The current studies explored consistent relationships among three productive vocabulary task scores and "infrequent lexical resource" in IELTS Writing Task 2 for a B2 proficiency level group, and both "fluency" and "infrequent lexical resource" in the IELTS speaking tasks for an A2 proficiency level group.

(iii) To replicate the four earlier studies (Chapters 3-6) but for receptive vocabulary knowledge. Future studies might investigate the extent to which multiple receptive vocabulary knowledge tasks detect change for different proficiency learners and how such change might relate to language receptive skills (i.e., reading and listening) in support of recent calls (e.g., Pellicer-Sanchez, 2018).

Appendices

Appendix 1 The Lexical Frequency Profile (LFP; Laufer & Nation, 1995; Chapter 2)

Composition Topic 1 (IELTS, 2019a)

The threat of nuclear weapons maintains world peace. Nuclear power provides cheap and clean energy.

The benefits of nuclear technology far outweigh the disadvantages.

To what extent do you agree or disagree?

Note. Adapted from https://www.ielts.org/for-test-takers/sample-test-questions

Composition Topic 2 (IELTS, 2019a)

The first car appeared on British roads in 1888. By the year 2020 there may be as many as 35 million vehicles on British roads. Alternative forms of transport should be encouraged and international laws introduced to control car ownership and use.

To what extent do you agree or disagree?

Note. Adapted from https://www.ielts.org/for-test-takers/sample-test-questions

Appendix 2 The Productive Vocabulary Level Test (PVLT; Laufer & Nation, 1999; Chapter 2)

Version A

Instruction: Complete the underlined words. The example has been done for you.

He was riding a bicycle.

The 2000-word level

1. I am glad we had this opp_____ to talk.
2. There are a doz_____eggs in the basket.
3. Every working person must pay income t_____.
4. The pirates buried the trea_____on a desert island.
5. Her beauty and cha_____ had a powerful effect on men.
6. La_____ of rain lead to a shortage of water in the city.
7. He takes cr_____and sugar in his coffee.
8. The rich man died and left all his we_____ to his son.
9. Pup_____ must hand in their papers by the end of the week.
10. This sweater is too tight. It needs to be stret_____.
11. Ann intro_____ her boyfriend to her mother.
12. Teenagers often adm_____ and worship pop singers.
13. If you blow up that balloon anymore it will bur_____.
14. In order to be accepted into the university, he had to impr_____ his grades.
15. The telegram was deli_____ to ours after it had been sent.
16. The differences were so sl_____ that they went unnoticed.
17. The dress you are wearing is lov_____.
18. He wasn't very popu_____ when he was a teenager, but he has many friends now.

The 3000-word level

1. He has a successful car_____ as a lawyer.
2. The thieves threw ac_____ in his face and made him blind.
3. To improve the country's economy, the government decided on economic ref_____.
4. She wore a beautiful green go_____ to the ball.
5. The government tried to protect the country's industry by reducing the imp_____ of cheap goods.
6. The children's games were funny at first, but finally got on the parents' ner_____.
7. The lawyer gave some wise coun_____ to his clients.
8. Many people in England mow the la_____ of their houses on Sunday morning.
9. The farmer sells the eggs that his he_____ lays.
10. Sudden noises at night sca_____ me a lot.
11. France was proc_____ a republic in the 18th century.
12. Many people are inj_____ in road accidents every year.
13. Suddenly, he was thru_____ into the dark room.
14. He perc_____ a light at the end of the tunnel.
15. Children are not independent. They are att_____ to their parents.
16. She showed off her sle_____ figure in a long narrow dress.
17. She has been changing partners often because she cannot have a sta_____ relationship with one person.
18. You must wear a bathing suit on a public beach. You're not allowed to be na_____.

The 5000-word level

1. Soldiers usually swear an oa _____ of loyalty to their country.
2. The voter placed the ball_____ in the box.
3. They keep their valuables in a vau_____ at the bank.
4. A bird perched at the window led_____.
5. The kitten is playing with a ball of ya_____.
6. The thieves have forced an ent_____ into the building.
7. The small hill was really a burial mou_____.
8. We decided to celebrate new year's e_____ together.
9. The solider was asked to choose between infantry and cav_____.
10. This is a complex problem which is difficult to compr_____.
11. The angry crowd sho_____ the prisoner as he was leaving the court.
12. Don't pay attention to this rude remark. Just ign_____ it.
13. The management held a secret meeting. The issues discussed were not disc_____ to the workers.
14. We could hear the sergeant bel_____ commands to the troops.
15. The boss got angry with the secretary and it took a lot of tact to soo_____ him.
16. We do not have adeq_____ information to make a decision.
17. She is not a child, but a mat_____ woman. She can make her own decisions.
18. The prisoner was put in soli_____ confinement.

The University Word List Level

1. There has been a recent tr_____ among prosperous families towards a smaller number of children.
2. The ar_____ of his office is 25 square meters.
3. Phil_____ examines the meaning of life.
4. According to the communist doc_____, workers should rule the world.
5. Spending many years together deepened their inti_____.
6. He usually read the sport sec_____ of the newspaper first.
7. Because of the doctors' strike the cli_____ is closed today.
8. There are several misprints on each page of this te_____.
9. The suspect had both opportunity and mot_____ to commit the murder.
10. They insp_____ all products before sending them out to stores.
11. A considerable amount of evidence was accum_____ during the investigation.
12. The victim's shirt was satu_____with blood.
13. He is irresponsible. You cannot re_____ on him for help.
14. It's impossible to eva_____ these results without knowing about the research methods that were used.
15. He finally att_____ a position of power in the company.
16. The story tells us about a crime and subs_____ punishment.
17. In a hom_____ class all students are of a similar proficiency.
18. The urge to survive is inh_____ in all creatures.

The 10000-word level

1. The baby is wet. Her dia_____ needs changing.
2. The prisoner was released on par_____.
3. Second year University students in the US are called soph_____.
4. Her favorite flowers were or_____.
5. The insect causes damage to plants by its toxic sec_____.
6. The evac_____ of the building saved many lives.
7. For many people, wealth is a prospect of unimaginable felic_____.
8. She found herself in a pred_____ without any hope for a solution.
9. The deac_____ helped with the care of the poor of the parish.
10. The hurricane whi_____ along the coast.
11. Some coal was still smol_____ among the ashes.
12. The dead bodies were muti_____ beyond recognition.
13. She was sitting on a balcony and bas_____ in the sun.
14. For years waves of invaders pill_____ towns along the coast.
15. The rescue attempt could not proceed quickly. It was imp_____ by bad weather.
16. I wouldn't hire him. He is unmotivated and indo_____.
17. Computers have made typewriters old-fashioned and obs_____.
18. Watch out for his wil_____ tricks.

Note. Reprinted from "A Vocabulary-Size Test of Controlled Productive Ability," by Laufer & Nation, 1999, *Journal of Language Testing*, 16(1), p. 46. Copyright 1999 by SAGE Journals.

Version B: The PVLT (Laufer & Nation, 1999)

Instruction: Complete the underlined words. The example has been done for you.

He was riding a bicycle.

The 2000-word level

1. It is the de_____ that counts, not the thought.
2. Plants receive water from the soil through their ro_____.
3. The nu _____was helping the doctor in the operation room.
4. Since he is unskilled, he earns low wa_____.
5. This year long sk_____ are fashionable again.
6. Laws are based upon the principle of jus_____.
7. He is walking on the ti_____ of his toes.
8. The mechanic had to replace the mo_____ of the car.
9. There is a co_____ of the original report in the file.
10. They had to cl_____ a steep mountain to reach the cabin.
11. The doctor ex_____ the patient thoroughly.
12. The house was su_____ by a big garden.
13. The railway con_____ London with its suburbs.
14. She wan_____ aimlessly in the street.
15. The organisers li_____ the number of participants to fifty.
16. This work is not up to your usu_____ standard.
17. They sat down to eat even though they were not hu_____.
18. You must have been very br_____ to participate in such a dangerous operation.

The 3000-word level

1. I live in a small apa_____ on the second floor.
2. The pro_____ of failing the test scared him.
3. Before writing the final version, the student wrote several dra_____.
4. It was a cold day. There was a ch_____ in the air.
5. The cart is pulled by an o_____.
6. Anthropologists study the struc_____ of ancient societies.
7. After two years in the Army, he received the rank of lieu_____.
8. The statue is made of mar_____.
9. Some aristocrats believed that blue blood flowed through their ve_____.
10. The secretary assi_____ the boss in organizing the course.
11. His beard was too long. He decided to tr_____ it.
12. People were whir_____ round on the dance floor.
13. He was on his knees, ple_____ for mercy.
14. You'll sn_____ that branch if you bend it too far.
15. I won't tell anybody. My lips are sea_____.
16. Crying is a nor_____ response to pain.
17. The Emperor of China was the supr_____ ruler of his country.
18. You must be awa_____ that very few jobs are available.

The 5000-word level

1. Some people find it difficult to become independent. Instead they prefer to be tied to their mother's ap_____ strings.
2. After finishing his degree, he entered upon a new ph_____ in his career.
3. The workmen cleaned up the me_____ before they left.
4. On Sunday, in his last se_____ in Church, the priest spoke against child abuse.
5. I saw them sitting on st_____ at the bar drinking beer.
6. Her favorite musical instrument was a tru_____.
7. The building is heated by a modern heating appa_____.
8. He received many com_____ on his dancing skill.
9. People manage to buy houses by raising a mor_____ from a bank.
10. At the bottom of a blackboard there is a le_____ for chalk.
11. After falling off his bicycle, the boy was covered with bru_____.
12. The child was holding a doll in her arms and hu_____ it.
13. We'll have to be inventive and de_____ a scheme for earning more money.
14. The picture looks nice; the colours bl_____ really well.
15. Nuts and vegetables are considered who_____ food.
16. The garden was full of fra_____ flowers.
17. Many people feel depressed and gl_____ about the future of the mankind.
18. He is so depressed that he is cont_____ suicide.

The University Word List Level

1. I've had my eyes tested and the optician says my vi_____ is good.
2. The anom_____ of his position is that he is the chairman of the committee, but isn't allowed to vote.
3. In their geography class, the children are doing a special pro_____ on North America.
4. In a free country, people can apply for any job. They should not be discriminated against on the basis of colour, age, or s_____ .
5. A true dem_____ should ensure equal rights and opportunities for all citizens.
6. The drug was introduced after medical res_____ indisputably proved its effectiveness.
7. These courses should be taken in seq_____, not simultaneously.
8. Despite his physical condition, his int_____ was unaffected.
9. Governments often cut budgets in times of financial cri_____.
10. The job offer sounded interesting at first. But when he realised what it would involve, his excitement subs_____ gradually.
11. Research ind_____ that men find it easier to give up smoking than women.
12. In a lecture, most of the talking is done by the lecturer. In a seminar, students are expected to part_____ in the discussion.
13. The airport is far away. If you want to ens_____ that you catch your plane, you have to leave early.
14. It's difficult to ass_____ a person's true knowledge by one or two tests.
15. The new manager's job was to res_____ the company to its former profitability.
16. Even though the student didn't do well on the midterm exam, he got the highest mark on the fi_____.
17. His decision to leave home was not well thought out. It was not based on rat_____ considerations.
18. The challenging job required a young, successful and dyn_____ candidate.

The 10000-word level

1. The new vic_____ was appointed by the bishop.
2. If your lips are sore, try lip sal_____, not medicine.
3. Much to his chag_____, he was not offered the job.
4. The actors exchanged ban_____ with reporters.
5. She wanted to marry nobility: a duke, a baron, or at least a vis_____.
6. The floor in the ballroom was a mos_____ of pastel colours.
7. She has contributed a lot of money to various charities. She is known for her generosity and bene_____.
8. This is an unusual singer with a range of three oct_____.
9. A thro_____ controls the flow of gas into an engine.
10. Anyone found loo_____ bombed houses and shops will be severly punished.
11. The crowd soon disp_____ when the police arrived.
12. The wounded man squi_____ on the floor in agony.
13. The dog crin_____ when it saw the snake.
14. He imme_____ himself in a hot bubbly bath forgetting all his troubles for a moment.
15. The approaching storm stam_____ the cattle into running wildly.
16. The problem is beginning to assume mam_____ proportions.
17. His vind_____ behaviour towards the thief was understandable.
18. He was arrested for illi_____ trading in drugs.

Note. Reprinted from "A Vocabulary-Size Test of Controlled Productive Ability," by Laufer & Nation, 1999, *Journal of Language Testing*, 16(1), p. 48–49. Copyright 1999 by SAGE Journals.

Version C: The PVLT (Laufer & Nation, 1999)

Instruction: Complete the underlined words. The example has been done for you.

He was riding a bicycle.

The 2000-word level

1. La ……. of rain led to a shortage of water in the city.
2. The rich man died and left all his we……. to his son.
3. Pup……. must hand in their papers by the end of the week.
4. This sweater is too tight. It needs to be stret……. .
5. If you blow up that balloon any more it will bur……. ..
6. In order to be accepted into the university, he had to impr……. his grades.
7. The differences were so sl……. that they went unnoticed.
8. The dress you're wearing is lov……. .
9. It is the de……. that counts, not the thought.
10. Plants receive water from the soil through their ro……. ..
11. The nu……. was helping the doctor in the operating room.
12. Since he is unskilled, he earns low wa……. .
13. This year long sk……. are fashionable again.
14. He is walking on the ti……. of his toes.
15. They had to cl……. a steep mountain to reach the cabin.
16. She wan……. aimlessly in the streets.
17. This work is not up to your usu……. standard.
18. They sat down to eat even though they were not hu……. .

The 3000-word level

1. She wore a beautiful green go……. to the ball.
2. Many people in England mow the la……. of their houses on Sunday morning.
3. The farmer sells the eggs that his he……. lays.
4. Sudden noises at night sca……. me a lot.
5. Many people are inj……. in road accidents every year.
6. Suddenly he was thru……. into the dark room.
7. She showed off her sle……. figure in a long narrow dress.
8. You must wear a bathing suit on a public beach. You're not allowed to bath na……. .
9. Before writing the final version, the student wrote several dra……. .
10. It was a cold day. There was a ch……. in the air.
11. The cart is pulled by an o……. .
12. His beard was too long. He decided to tr……. it.
13. People were whir……. around on the dance floor.
14. You'll sn……. that branch if you bend it too far.
15. I won't tell anybody. My lips are sea……. .
16. You must be aw……. that very few jobs are available.
17. After two years in the Army, he received the rank of lieu……. .
18. The pro……. of failing the test scared him.

The 5000-word level

1. Soldiers usually swear an oa ……. of loyalty to their country.
2. The voter placed the ball ……. in the box.
3. They keep their valuables in a vau ……. at the bank.
4. The kitten is playing with a ball of ya ……. .
5. We decided to celebrate New Year's E ……. together.
6. We could hear the sergeant bel ……. commands to the troops.
7. The boss got angry with the secretary and it took a lot of tact to soo……. him.
8. Some people find it difficult to become independent. They prefer to be tied to their mother's ap……. strings.
9. The workmen cleaned up the me……. before they left.
10. I saw them sitting on st……. at the bar drinking beer.
11. People manage to buy houses by raising a mor……. from a bank.
12. At the bottom of the blackboard there is a le……. for chalk.
13. After falling off his bicycle, the boy was covered with bru……. .
14. The child was holding a doll in her arms and hu……. it.
15. The picture looks nice; the colours bl……. really well.
16. Nuts and vegetables are considered who……. food.
17. Many gardens are full of fra……. flowers.
18. Many people feel depressed and gl……. about the future of mankind.

The University Word List Level

1. There has been a recent tr…….among prosperous families towards a smaller number of children.
2. The ar…….of his office is 25 square meters.
3. Phil…….examines the meaning of life.
4. According to the communist doc……., workers should rule the world.
5. Spending many years together deepened their inti……. .
6. He usually read the sport sec……. of the newspaper first.
7. Because of the doctors' strike the cli……. is closed today.
8. There are several misprints on each page of this te……. .
9. The suspect had both opportunity and mot…….to commit the murder.
10. They insp…….all products before sending them out to stores.
11. A considerable amount of evidence was accum…….during the investigation.
12. The victim's shirt was satu…….with blood.
13. He is irresponsible. You cannot re…….on him for help.
14. It's impossible to eva…….these results without knowing about the research methods that were used.
15. He finally att…….a position of power in the company.
16. The story tells us about a crime and subs……. punishment.
17. In a hom……. class all students are of a similar proficiency.
18. The urge to survive is inh…….in all creatures.

The 10000-word level

1. The baby is wet. Her dia……. needs changing.
2. Second year university students in the US are called soph……. .
3. The deac……. helped with the care of the poor of the parish.
4. The hurricane whi……. along the coast.
5. Some coal was still smol……. among the ashes.
6. She was sitting on a balcony and bas……. in the sun.
7. Computers have made typewriters old-fashioned and obs……. .
8. Watch out for his wil……. tricks.
9. If your lips are sore, try lip sal……. , not medicine.
10. The new vic……. was appointed by the bishop.
11. The actors exchanged ban……. with the reporters.
12. A thro……. controls the flow of gas into an engine.
13. Anyone found loo……. bombed houses and shops will be severely punished.
14. The wounded man squi……. on the floor in agony.
15. The dog crin……. when it saw the snake.
16. The approaching storm stam……. the cattle into running wildly.
17. The problem is beginning to assume mam……. proportions.
18. The rescue attempt could not proceed quickly. It was imp……. by bad weather**.**

Note. Adapted from "A Vocabulary-Size Test of Controlled Productive Ability," by Laufer & Nation, 1999, *Journal of Language Testing*, 16(1), p. 46–51. Copyright 1999 by SAGE Journals.

Appendix 3 Lex30 (Meara & Fitzpatrick, 2000; Chapter 2)

Version A

Instruction: Write down the first four (English) words you think of when you read each word in the list.

1.	attack				
2.	board				
3.	close				
4.	cloth				
5.	dig				
6.	dirty				
7.	disease				
8.	experience				
9.	fruit				
10.	furniture				
11.	habit				
12.	hold				
13.	hope				
14.	kick				
15.	map				
16.	obey				
17.	pot				
18.	potato				
19.	real				
20.	rest				
21.	rice				
22.	science				
23.	seat				
24.	spell				
25.	substance				
26.	stupid				
27.	television				
28.	tooth				
29.	trade				
30.	window				

Note. Reprinted from "Lex30: An Improved Method of Assessing Productive Vocabulary in an L2," by Meara & Fitzpatrick, 2000, *System Journal*, 28(1), p. 28–29. Copyright 2000 by Elsevier.

Version B: Lex30 (Meara & Fitzpatrick, 2000)

Instruction: Write down the first four (English) words you think of when you read each word in the list.

1.	away				
2.	blow				
3.	brush				
4.	chance				
5.	common				
6.	dance				
7.	district				
8.	ever				
9.	famous				
10.	flag				
11.	get				
12.	head				
13.	insect				
14.	knee				
15.	list				
16.	mat				
17.	mountain				
18.	oil				
19.	pattern				
20.	policeman				
21.	public				
22.	religion				
23.	secret				
24.	shirt				
25.	sorry				
26.	smell				
27.	spirit				
28.	surprise				
29.	telephone				
30.	tool				

Note. Adapted from "Lex30: An Improved Method of Assessing Productive Vocabulary in an L2," by Meara & Fitzpatrick, 2000, *System Journal*, 28(1), p. 28–29. Copyright 2000 by Elsevier.

Version C: Lex30

Instruction: Write down the first four (English) words you think of when you read each word in the list.

1.	find				
2.	fish				
3.	walk				
4.	water				
5.	sleep				
6.	cold				
7.	bird				
8.	light				
9.	sea				
10.	paper				
11.	friend				
12.	tell				
13.	eye				
14.	jump				
15.	book				
16.	think				
17.	glass				
18.	music				
19.	fire				
20.	give				
21.	money				
22.	car				
23.	army				
24.	slow				
25.	train				
26.	cry				
27.	sun				
28.	end				
29.	bed				
30.	door				

Note. The third (C) version of the task was created based on the original task criteria. The original task is from "Lex30: An Improved Method of Assessing Productive Vocabulary in an L2," by Meara & Fitzpatrick, 2000, *System Journal*, 28(1), p. 28–29. Copyright 2000 by Elsevier.

Appendix 4 G_Lex (Fitzpatrick & Clenton, 2017; Chapter 2)

Version A

Instruction: Write down five different words that might fit into each gap. The gaps are suitable for nouns, adjectives, and verbs in equal measure (eight sentences each).

1. She loved to _____ over the phone.					
2. When I feel sad, I always go to the_____.					
3. They think car-racing is_____.					
4. His colleague wanted to _____ the report.					
5. My favourite _____ is football.					
6. She looked _____ when she saw her friends.					
7. He couldn't _____ the car.					
8. With a fire in my house, I would save my ____.					
9. Many people feel _about the environment.					
10. The parents _____ the children.					
11. He was happy with his _____.					
12. He didn't think her teacher was _____ at all.					
13. She always wanted to _____ after a busy day at work.					
14. She sent _____to her mother.					
15. The weather looked ____ before the game.					
16. He wanted to _____ the letter.					
17. She was excited about _____.					
18. The girls thought the rock concert was____.					
19. He took the chance to _____ the president.					
20. He gave his boss _____.					
21. At the funeral the family felt ___.					
22. He always _____his breakfast.					
23. She put the food in the _____.					
24. She was always ____ to those who needed help.					

Note. Adapted from "Making Sense of Learner Performance on Tests of Productive Vocabulary Knowledge," by Fitzpatrick & Clenton, 2017, *TESOL Quarterly Journal*, 51(4), p. 856.

Version B: G_Lex

Instruction: Write down five different words that might fit into each gap. The gaps are suitable for nouns, adjectives, and verbs in equal measure (eight sentences each).

He liked to _____ in his free time.					
When I feel happy, I always go to the_____.					
They think tennis is_____.					
He wanted to _____ the homework.					
My best _____was in Japan.					
She felt _____ when she met her friends.					
She could _____ the bicycle.					
On my next trip I would like to buy ____.					
My parents feel ____ about my future plans.					
The teachers _____ the students.					
He was sad about his _____.					
He thought his friend was _____.					
She wanted to _____ next year.					
She bought _____for her father.					
The players looked ____ before the game.					
He wanted to _____ the email.					
She was nervous about her _____.					
They thought the movie was____.					
He tried to _____ his boss.					
She gave her friend _____.					
At the wedding party, the family felt ___.					
He always _____his keys.					
She put her new toy on the _____.					
They are_____ people.					

Note. The second (B) version of the task was created based on the original task criteria. The original task is from "Making Sense of Learner Performance on Tests of Productive Vocabulary Knowledge," by Fitzpatrick & Clenton, 2017, *TESOL Quarterly Journal*, 51(4), p. 856. Copyright 2017 by TESOL International Association.

Version C: G_Lex

Instruction: Write down five different words that might fit into each gap. The gaps are suitable for nouns, adjectives, and verbs in equal measure (eight sentences each).

1. He tried to _____ during his summer vacation.					
2. When I feel angry, I always go to the_____.					
3. They think football is_____.					
4. She wanted to _____ the project.					
5. My best _____is orange.					
6. She felt _____ when she received her test score.					
7. She couldn't _____ the house.					
8. She should include more ____ in her next report.					
9. My friends feel ____ about my new car.					
10. The government _____ the people.					
11. He was surprised about his _____.					
12. He thought his parents were _____.					
13. She wanted to _____ her life.					
14. She sent _____to her boss.					
15. We looked ____ before the game.					
16. He wanted to _____ the message.					
17. She was happy about her _____.					
18. They thought the basketball game was____.					
19. He tried to _____ his teacher.					
20. She gave her mother _____.					
21. At the graduation party, the family felt ___.					
22. She always _____her bag.					
23. Last night, I had my worst _____.					
24. They are_____ players.					

Note. The Third (C) version of the task was created based on the original task criteria. The original task is from "Making Sense of Learner Performance on Tests of Productive Vocabulary Knowledge," by Fitzpatrick & Clenton, 2017, *TESOL Quarterly Journal*, 51(4), p. 856.

Appendix 5 Samples of A2 Participants' Responses to Lex30 at 0 and 3 Months

(Chapter 4)

Lex30 for A2 Participants (Test Time 1)

Lex30

(Meara & Fitzpatrick, 2000)

Name (in English):..... Code: 8L5

Date: . Class:. Day:. Period:

Time: 15 minutes

Instruction: Write down the first four (English) words you think of when you read each word in the list.

1.	away	house	run		
2.	blow	wing			
3.	brush	up	tooth	car	bike
4.	chance	again	game	try	positive
5.	common	dance	word	story	topic
6.	dance	together	enjoy	sport	
7.	district	ban			
8.	ever	since	time	yet	
9.	famous	food	clothes	shoes	paper
10.	flag	rise	raise	country	mark
11.	get	pen	eraser	wallet	phone
12.	head	important	complex		
13.	insect	unlike	summer	youngstar	
14.	knee	body	break		
15.	list	word			
16.	mat	wet			
17.	mountain	climb	seem	cold	attach
18.	oil	delicious	convenience	cooking	dish
19.	pattern				
20.	policeman	important	scared	smart	
21.	public	city	ocean	park	toilet
22.	religion	people	blieve		
23.	secret	garden	subject		
24.	shirt	dirty	clean	white	black
25.	sorry	face	negative		
26.	smell	good	sweet	bad	flower
27.	spirit	strong	heart		
28.	surprise	face	impression	UFO	god
29.	telephone	ring	talk		
30.	tool	goods	technology		

Lex30 for A2 Participants (Test Time 2)

Lex30

Adapted from Meara & Fitzpatrick (2000)

Name (in English): Code: 8 L5

Date: Class: Day: Period:

Time: 15 minutes

Instruction: Write down the first four (English) words you think of when you read each word in the list.

1.	find	money	watch	bag	phone
2.	sun	hot	happy	warm	Summer
3.	walk	foot	shoes	sunny	walking
4.	water	drink	moush	thirsty	clean
5.	sleep	eye	bed	room	house
6.	cold	winter	snow	white	cloud
7.	bird	sky	blue	fly	chicken
8.	give	present	Christmas	Birthday	happiness
9.	sea	huge	blue	island	Japan
10.	paper	write	white	pencil	eraser
11.	book	read	smart	study	exam
12.	tell	story	episode	talk	people
13.	eye	contact	see	two	shame
14.	jump	foot	happy	fly	ground
15.	friend	partner	girl	talking	believe
16.	think	future	thing	bother	quiet
17.	glass	clean	water	drink	talk
18.	music	ear	walkman	phone	personality
19.	fish	swim	water	river	sea
20.	end	finish	goal	depress.	running
21.	fire	house	employ	red	work
22.	car	drive	road	incident	accident
23.	army	war	running	quiet	smart
24.	slow	quick	quiet	walk	stop
25.	train	transform	people	load	signal
26.	cry	face	tear	child	loud
27.	fish	breath	water	eat	fresh
28.	light	dark	shine	future	positive
29.	money	working	hard	decrease	pleasure
30.	door	house	border	furniture	wood

Appendix 6 Samples of A2 Participants' Responses to G_Lex at 0 and 3 Months

(Chapter 4)

G_Lex for A2 Participants (Test Time 1)

G-Lex

Adapted from Fitzpatrick & Clenton (2017)

Name (in English): Code: 18 L 3

Date Class. Day: Period:

Time: 15 minutes

Instruction: Write down five different words that might fit into each gap. The gaps are suitable for nouns, adjectives, and verbs in equal measure (eight sentences each).

1. He liked to _____ in his free time.	play	sleep	walk		
2. When I feel happy I always go to the_____.	school	stadiam	court		
3. They think tennis is_____.	tired	fun			
4. He wanted to _____ the homework.	miss	do	rent		
5. My best _____was in Japan.	country				
6. She felt _____ when she met her friends.	glad	sad			
7. She could _____ the bicycle.	buy	rent	borrow		
8. On my next trip I would like to buy ___.	land	house	nuts		
9. My parents feel ____ about my future plans.	mud	nervos	cry		
10. The teachers _____ the students.	scold				
11. He was sad about his _____.	salary	dream			
12. He thought his friend was _____.	employed	fired			
13. She wanted to _____ next year.	marry				
14. She bought _____for her father.	car	golfclab			
15. The players looked ____ before the game.	nice	tired	bad		
16. He wanted to _____ the email.	stop	send			
17. She was nervous about her _____.	dream	test	friend		
18. They thought the movie was____.					
19. He tried to _____ his boss.	met	talk to			
20. She gave her friend _____.	shirt	watch			
21. At the wedding party, the family felt ___.	cry	sad			
22. He always _____his keys.	miss				
23. She put her new toy on the _____.	table	seat	floor		
24. They are_____ people..	helpful	useful			

G_Lex for A2 Participants (Test Time 2)

G-Lex

Adapted from Fitzpatrick & Clenton (2017)

Name (in English): Code: 18L3

Date: Class: Day: Period:

Time: 15 minutes

Instruction: Write down five different words that might fit into each gap. The gaps are suitable for nouns, adjectives, and verbs in equal measure (eight sentences each).

1. He tried to _____ during his summer vacation.	study	experiment	swim	tennis	go back
2. When I feel angry I always go to the_____.	bed	beach	mountain	lake	park
3. They think football is_____.	fun	dangerous	hard	speedy	hot
4. She wanted to _____ the project.	plan	do	get	forget	
5. My best _____is orange.	fruits	colors			
6. She felt _____ when she received her test score.	angry	happy	sad	surprised	sick
7. She couldn't _____ the house.	get	buy	perchase	build	destroy
8. She should include more ____ in her next report.	words	interesting			
9. My friends feel ____ about my new car.	happy	sad	great	angry	surprised
10. The government _____ the people.	see				
11. He was surprised about his _____.	scores	cars	bags	shoes	organs
12. He thought his parents were _____.	died	left	angry	waiting	
13. She wanted to _____ her life.	get	break			
14. She sent _____to her boss.	e-mail	letter			
15. We looked ____ before the game.	fine	happy	sad		
16. He wanted to _____ the message.	return	write	tear	read	vanish
17. She was happy about her _____.	money	graduation			
18. They thought the basketball game was____.	fun	dangerous	speedy		
19. He tried to _____ his teacher.	complant				
20. She gave her mother _____.	watch	bags	shoes		
21. At the graduation party, the family felt ___.	happy				
22. She always _____her bag.	get				
23. Last night, I had my worst _____.	sleep	day	drink		
24. They are_____ players.	tennis	baseball	basketball		

Appendix 7 Samples of A2 Participants' Responses to the PVLT at 0 and 3 Months

(Chapter 4)

PVLT for A2 Participants: 2,000- and 3,000-Word Levels (Test Time 1)

Productive Vocabulary Level Test
(PVLT; Laufer and Nation, 1999)

Name (in English): Code: 813

Date: Class: Day: Period:

Instruction: Complete the underlined words. The example has been done for you.
He was riding a bicycle.

The 2000-word level

1. It is the de_____ that counts, not the thought.
2. Plants receive water from the soil through their ro_____.
3. The nu rs was helping the doctor in the operation room.
4. Since he is unskilled, he earns low wa_____.
5. This year long sk_____ are fashionable again
6. Laws are based upon the principle of jus tole.
7. He is walking on the ti_____ of his toes.
8. The mechanic had to replace the mo nitor of the car.
9. There is a co mputer of the original report in the file.
10. They had to cl_____ a steep mountain to reach the cabin.
11. The doctor ex_____ the patient thoroughly.
12. The house was su___ ___ by a big garden.
13. The railway con bine London with its suburbs.
14. She wan_____ aimlessly in the street.
15. The organisers li_____ the number of participants to fifty.
16. This work is not up to your usu al ___ standard.
17. They sat down to eat even though they were not hu_____.
18. You must have been very br_____ to participate in such a dangerous operation.

The 3000-word level

1. I live in a small apa rt on the second floor.
2. The pro file of failing the test scared him.
3. Before writing the final version, the student wrote several dra_____.
4. It was a cold day. There was a ch_____ in the air.
5. The cart is pulled by an o_____.
6. Anthropologists study the struc tura of ancient societies.
7. After two years in the Army, he received the rank of lieu_____.
8. The statue is made of mar_____.
9. Some aristocrats believed that blue blood flowed through their ve_____
10. The secretary assi stance the boss in organizing the course.
11. His beard was too long. He decided to tr_____ it.
12. People were whir_____ round on the dance floor.
13. He was on his knees, ple_____ for mercy.
14. You'll sn_____ that branch if you bend it too far.
15. I won't tell anybody. My lips are sea_____.
16. Crying is a nor_____ response to pain.
17. The Emperor of China was the supreme ruler of his country.
18. You must be awa_____ that very few jobs are available.

PVLT for A2 Participants: 5,000- and UWL-Word Levels (Test Time 1)

The 5000-word level

1. Some people find it difficult to become independent. Instead they prefer to be tied to their mother's ap_____ strings.
2. After finishing his degree, he entered upon a new ph_____ in his career.
3. The workmen cleaned up the me_____ before they left.
4. On Sunday, in his last se_____ in Church, the priest spoke against child abuse.
5. I saw them sitting on st_____ at the bar drinking beer.
6. Her favorite musical instrument was a tru_____.
7. The building is heated by a modern heating appa_____.
8. He received many com_____ on his dancing skill.
9. People manage to buy houses by raising a mor_____ from a bank
10. At the bottom of a blackboard there is a le_____ for chalk.
11. After falling off his bicycle, the boy was covered with bru_____.
12. The child was holding a doll in her arms and hu_____ it.
13. We'll have to be inventive and de_____ a scheme for earning more money.
14. The picture looks nice; the colours bl_____ really well.
15. Nuts and vegetables are considered who_____ food.
16. The garden was full of fra_____ flowers.
17. Many people feel depressed and gl_____ about the future of the mankind.
18. He is so depressed that he is cont_____ suicide.

The University Word List Level

1. I've had my eyes tested and the optician says my vi_____ is good.
2. The anom_____ of his position is that he is the chairman of the committee, but isn't allowed to vote.
3. In their geography class, the children are doing a special pro_____ on North America.
4. In a free country, people can apply for any job. They should not be discriminated against on the basis of colour, age, or s_____.
5. A true dem_____ should ensure equal rights and opportunities for all citizens.
6. The drug was introduced after medical res_____ indisputably proved its effectiveness.
7. These courses should be taken in seq_____, not simultaneously.
8. Despite his physical condition, his int_____ was unaffected.
9. Governments often cut budgets in times of financial cri_____.
10. The job offer sounded interesting at first. But when he realised what it would involve, his excitement subs_____ gradually.
11. Research ind_____ that men find it easier to give up smoking than women.
12. In a lecture, most of the talking is done by the lecturer. In a seminar, students are expected to part_____ in the discussion.
13. The airport is far away. If you want to ens_____ that you catch your plane, you have to leave early.
14. It's difficult to ass_____ a person's true knowledge by one or two tests.
15. The new manager's job was to res_____ the company to its former profitability.
16. Even though the student didn't do well on the midterm exam, he got the highest mark on the fi_____.
17. His decision to leave home was not well thought out. It was not based on rat_____ considerations.
18. The challenging job required a young, successful and dyn_____ candidate.

PVLT for A2 Participants: 10,000-Word Levels (Test Time 1)

The 10000-word level

1. The new vic_____ was appointed by the bishop.
2. If your lips are sore, try lip sal_____, not medicine.
3. Much to his chag_____, he was not offered the job.
4. The actors exchanged ban_____ with reporters.
5. She wanted to marry nobility: a duke, a baron, or at least a vis_____.
6. The floor in the ballroom was a mos_____ of pastel colours.
7. She has contributed a lot of money to various charities. She is known for her generosity and bene_____.
8. This is an unusual singer with a range of three oct_____.
9. A thro_____ controls the flow of gas into an engine.
10. Anyone found loo_____ bombed houses and shops will be severly punished.
11. The crowd soon disp_____ when the police arrived.
12. The wounded man squi_____ on the floor in agony.
13. The dog crin when it saw the snake.
14. He imme_____ himself in a hot bubbly bath forgetting all his troubles for a moment.
15. The approaching storm stam_____ the cattle into running wildly.
16. The problem is beginning to assume mam_____ proportions.
17. His vind_____ behaviour towards the thief was understandable.
18. He was arrested for illi_____ trading in drugs.

PVLT for A2 Participants: 2,000- and 3,000-Word Levels (Test Time 2)

Productive Vocabulary Level Test
(PVLT; Laufer and Nation, 1999)

Name (in English) Code:.....8L3..........................

Date: Class: Day: Period:

Instruction: Complete the underlined words. The example has been done for you.
He was riding a bicycle.

The 2000-word level

1. La .... of rain led to a shortage of water in the city.
2. The rich man died and left all his we............. to his son.
3. Pupular.... must hand in their papers by the end of the week.
4. This sweater is too tight. It needs to be stret.c.h.. .... .
5. If you blow up that balloon any more it will bur............. ..
6. In order to be accepted into the university, he had to impression. his grades.
7. The differences were so sl............. that they went unnoticed.
8. The dress you're wearing is love.l.y. ....
9. It is the de............. that counts, not the thought.
10. Plants receive water from the soil through their ro............. ..
11. The nu.r.s... .... was helping the doctor in the operating room.
12. Since he is unskilled, he earns low wa............. .
13. This year long sk.irt... ..... are fashionable again.
14. He is walking on the ti............. of his toes.
15. They had to cl.i.m.. ..... a steep mountain to reach the cabin.
16. She wan............. aimlessly in the streets.
17. This work is not up to your usu............. standard.
18. They sat down to eat even though they were not hu............. .

The 3000-word level

1. She wore a beautiful green go .... to the ball.
2. Many people in England mow the la............. of their houses on Sunday morning.
3. The farmer sells the eggs that his he............. lays.
4. Sudden noises at night sca.r.y.. ..... me a lot.
5. Many people are inj.ur.e... .... in road accidents every year.
6. Suddenly he was thru............. into the dark room.
7. She showed off her sle............. figure in a long narrow dress.
8. You must wear a bathing suit on a public beach. You're not allowed to bath na............. .
9. Before writing the final version, the student wrote several dra............. .
10. It was a cold day. There was a ch............. in the air.
11. The cart is pulled by an o............. .
12. His beard was too long. He decided to tr............. it.
13. People were whir............. around on the dance floor.
14. You'll sn............. that branch if you bend it too far.
15. I won't tell anybody. My lips are sea............. .
16. You must be aw............. that very few jobs are available.
17. After two years in the Army, he received the rank of lieu............. .
18. The pro............. of failing the test scared him.

PVLT for A2 Participants: 5,000- and UWL-Word Levels (Test Time 2)

The 5000-word level

1. Soldiers usually swear an oa …… of loyalty to their country.
2. The voter placed the ball …… in the box.
3. They keep their valuables in a vau …… at the bank.
4. The kitten is playing with a ball of ya …… .
5. We decided to celebrate New Year's E …… together.
6. We could hear the sergeant bel …… commands to the troops.
7. The boss got angry with the secretary and it took a lot of tact to soo…… him.
8. Some people find it difficult to become independent. They prefer to be tied to their mother's ap……. strings.
9. The workmen cleaned up the me…… before they left.
10. I saw them sitting on st…… at the bar drinking beer.
11. People manage to buy houses by raising a mor…… from a bank.
12. At the bottom of the blackboard there is a le…… for chalk.
13. After falling off his bicycle, the boy was covered with bru…… .
14. The child was holding a doll in her arms and hu…… it.
15. The picture looks nice; the colours bl…… really well.
16. Nuts and vegetables are considered who…… food.
17. Many gardens are full of fra…… flowers.
18. Many people feel depressed and gl…… about the future of mankind.

The University Word List Level

1. There has been a recent tr……among prosperous families towards a smaller number of children.
2. The ar……of his office is 25 square meters.
3. Phil……examines the meaning of life.
4. According to the communist doc……., workers should rule the world.
5. Spending many years together deepened their inti…… .
6. He usually read the sport sec…… of the newspaper first.
7. Because of the doctors' strike the cli…… is closed today.
8. There are several misprints on each page of this te…… .
9. The suspect had both opportunity and mot……to commit the murder.
10. They insp……all products before sending them out to stores.
11. A considerable amount of evidence was accum……during the investigation.
12. The victim's shirt was satu……with blood.
13. He is irresponsible. You can not re……on him for help.
14. It's impossible to eva……these results without knowing about the research methods that were used.
15. He finally att……a position of power in the company.
16. The story tells us about a crime and subs…… punishment.
17. In a hom…… class all students are of a similar proficiency.
18. The urge to survive is inh……in all creatures.

PVLT for A2 Participants:10,000-Word Levels (Test Time 2)

The 10000-word level

1. The baby is wet. Her dia..... needs changing.
2. Second year university students in the US are called soph..... .
3. The deac..... helped with the care of the poor of the parish.
4. The hurricane whi..... along the coast.
5. Some coal was still smol..... among the ashes.
6. She was sitting on a balcony and bas..... in the sun.
7. Computers have made typewriters old-fashioned and obs..... .
8. Watch out for his wil..... tricks.
9. If your lips are sore, try lip sal..... , not medicine.
10. The new vic..... was appointed by the bishop.
11. The actors exchanged ban..... with the reporters.
12. A thro..... controls the flow of gas into an engine.
13. Anyone found loo..... bombed houses and shops will be severely punished.
14. The wounded man squi..... on the floor in agony.
15. The dog crin..... when it saw the snake.
16. The approaching storm stam..... the cattle into running wildly.
17. The problem is beginning to assume mam..... proportions.
18. The rescue attempt could not proceed quickly. It was imp..... by bad weather.

Appendix 8 Samples of B2 Participants' Responses to the Lex30 at 0 and 3 Months

(Chapter 5)

Lex30 for B2 Participants (Test Time 1)

CODE: MONT-T1-27

Time: 15 minutes

Instruction: Write down the first four (English) words you think of when you read each word in the list.

1.	attack	bomb	agression	battle	castle
2.	board	class	black	write	
3.	close	open	store	night	buy
4.	cloth	friendship	love	sitting	
5.	dig	find	looking for	treasure	plant
6.	dirty	clean	housework	dust	
7.	disease	hospital	casualties	medecine	doctors
8.	experience	adiventure	improove	discover	try
9.	fruit	tree	peach	summer	strawberry
10.	furniture	table	chair	decoration	house
11.	habit	usually	bring		
12.	hold	support	bring	hand	friend
13.	hope	believe	faith	destiny	pray
14.	kick	leg	hit	fight	sports
15.	map	searching	direction	drive	travel
16.	obey	order	authority	leader	totalitarism
17.	pot				
18.	potato	fast food	eat	sweet	french fries
19.	real	concrete	life	dream	unbelievable
20.	rest	map	sleep	lay down	bed
21.	rice	thai	cook	risotto	food
22.	science	experiment	study	searchers	biology
23.	seat	car	belt	place	
24.	spell	words	letters	learn	write
25.	substance	liquide	strange	scary	texture
26.	stupid	dumb	stubborn	action	maturity
27.	television	shows	Netflix	entertainment	remote
28.	tooth	brush	toothpaste	white	smile
29.	trade	exchange	boat	buy	export
30.	window	outside	garden	glass	house

Lex30 for B2 Participants (Test Time 2)

Instruction: Write down the first four (English) words you think of when you read each word in the list.

1.	away	discover	Travel	foreign	leave
2.	blow	wind	mouth	candles	birthday
3.	brush	horse	pomey pony?	hair curls	curls ?
4.	chance	opportunity	luck	games	allowed
5.	common	casual	everyday	usual	habits
6.	dance	moves	music	enjoy	party
7.	district	neighborhood	New York	separation	street
8.	ever	forever	always	never	once
9.	famous	celebrity	singer	artist	actor
10.	flag	country	colors	pride	identity
11.	get	sease	catch	keep	finally
12.	head	face	body	top	hierarchy
13.	insect	bee	fly	wings	mosquito
14.	knee	leg	articulation	down	body
15.	list	write	count	summary	do
16.	mat				
17.	mountain	hill	landscape	nature	climb
18.	oil	cook	olives	Italy	tomato
19.	pattern	Scheme	motif	clothes	dots
20.	policeman	justice	work	car	arrestation
21.	public	theater	concert	crowd	common
22.	religion	christian	holy	choice	education
23.	secret	hide	dissimulate	share	personal
24.	shirt	dress	clothes	style	buttons
25.	sorry	apologize	excuse	guilty	relationship
26.	smell	food	grass	perfume	good
27.	spirit	holy	angel	mind	consciousness
28.	surprise	friends	birthday	present	happy
29.	telephone	call	mobile	technology	number
30.	tool	make	garden	hammer	fix

Appendix 9 Samples of B2 Participants' Responses to the G_Lex at 0 and 3 Months (Chapter 5)

G_Lex for B2 Participants (Test Time 1)

MONT-T1-28

Instruction: Write down five different words that might fit into each gap. The gaps are suitable for nouns, adjectives, and verbs in equal measure (eight sentences each).

1. She loved to ~~Spak~~ over the phone.	speak	play	call	shout	throw
2. When I feel sad I always go to the ~~bed~~.	bed	kitchen	bathroom	room	garden
3. They think car-racing is ~~cool~~.	cool	dangerous	outdated	useless	hard
4. His colleague wanted to ~~See~~ the report.	See	read	study	rework	finish
5. My favourite_____is football.	Sport	game	playing	pleasure	entertainment
6. She looked_____when she saw her friends.	concerned	happy	relieved	over	down
7. He couldn't_____the car.	drive	get	buy	sell	try
8. With a fire in my house I would save my____.	money	works	documents	family	animal
9. Many people feel____about the environment.	concerned	angry	great	indifferent	pressed
10. The parents_____the children.	loved	hated	sent	received	beat
11. He was happy with his_____.	wife	family	job	work	life
12. He didn't think her teacher was_____at all.	hot	good	bad	competent	special
13. She always wanted to_____after a busy day at work.	sleep	eat	rest	nap	go out
14. She sent_____to her mother.	money	cards	wishes	threats	children
15. The weather looked_____before the game.	ok	good	bad	fine	great
16. He wanted to_____the letter.	write	read	send	finish	burn
17. She was excited about_____.	me	working	resting	eating	moving
18. The girls thought the rock concert was____.	great	fine	good	weird	bad
19. He took the chance to_____the president.	become	meet	fire	elect	see
20. He gave his boss_____.	time	warnings	reports	pride	difficulties
21. At the funeral the family felt___.	sad	destroyed	overwhelmed	least	support
22. He always_____his breakfast.	skip	eats	forgets	takes	vomits
23. She put the food in the_____.	fridge	box	trash can	sink	mouth
24. She was always____to those who needed help.	~~great~~ helpfull	~~fde~~ supportive	~~helpful~~ mean	open	ready

G_Lex for B2 Participants (Test Time 2)

Instruction: Write down five different words that might fit into each gap. The gaps are suitable for nouns, adjectives, and verbs in equal measure (eight sentences each).

1. He liked to _____ in his free time.	rest	dance	play	sleep	
2. When I feel happy I always go to the_____.	bathroom	kitchen	bedroom	internet	
3. They think tennis is_____.	bad	fun	a sport	a joke	hard
4. He wanted to _____ the homework.	do	finish	skip	read	get
5. My best _____was in Japan.	vacation	life	friend	dream	trip
6. She felt _____ when she met her friends.	happy	relived	good		
7. She could _____ the bicycle.	ride	find	hide	break	use
8. On my next trip I would like to buy ____.	food	souvenirs	clothes	tickets	gifts
9. My parents feel ____ about my future plans.	uneven	weird	unconvinced	good	happy
10. The teachers _____ the students.	love	hate	torture	were	or
11. He was sad about his _____.	what?	life	test	failures	loses
12. He thought his friend was _____.	cool	smart	weird	popular	fine
13. She wanted to _____ next year.	move	die	come back	leave	
14. She bought _____for her father.	gifts	medeine	food	baguettes	something
15. The players looked ____ before the game.	tired	motivated	good	in-shape	ready
16. He wanted to _____ the email.	send	write	find	delete	read
17. She was nervous about her _____.	marriage	test	pregnancy	relationship	future
18. They thought the movie was____.	good	bad	new	long	perfect
19. He tried to _____ his boss.	fuck	kill	punch	please	help
20. She gave her friend _____.	love	support	gifts	help	advice
21. At the wedding party, the family felt ___.	loved	happy	ready	unsure	
22. He always _____his keys.	loses	forgets	has	gets	finds
23. She put her new toy on the _____.	toilet	bed	box	side	fridge
24. They are_____ people..	the	my	mean	good	weird

Appendix 10 Samples of B2 Participants' Responses to the PVLT at 0 and 3 Months (Chapter 5)

PVLT for B2 Participants: 2,000- and 3,000-Word Levels (Test Time 1)

MONT-T1-27

CODE :

NAME: (first) _____ (Surname) _____

Instruction: Complete the underlined words. The example has been done for you.

EXAMPLE: He was riding a bic_____. → He was riding a bicycle.

The 2000-word level

1. I am glad we had this opp ortunity to talk.
2. There are a doz en eggs in the basket.
3. Every working person must pay income t axes .
4. The pirates buried the trea sure on a desert island.
5. Her beauty and cha rm had a powerful effect on men.
6. La ck of rain lead to a shortage of water in the city.
7. He takes cr eam and sugar in his coffee.
8. The rich man died and left all his we alth to his son.
9. Pup ils must hand in their papers by the end of the week.
10. This sweater is too tight. It needs to be stret ched .
11. Ann intro duces her boyfriend to her mother.
12. Teenagers often adm ire and worship pop singers.
13. If you blow up that balloon anymore it will bur st .
14. In order to be accepted into the university, he had to impr ove his grades.
15. The telegram was deli vered to ours after it had been sent.
16. The differences were so sl ight that they went unnoticed.
17. The dress you are wearing is lov ely .
18. He wasn't very popu lar when he was a teenager, but he has many friends now.

The 3000-word level

1. He has a successful car eer as a lawyer.
2. The thieves threw ac id in his face and made him blind.
3. To improve the country's economy, the government decided on economic ref orms .
4. She wore a beautiful green go wn to the ball.
5. The government tried to protect the country's industry by reducing the imp act . of cheap goods.
6. The children's games were funny at first, but finally got on the parents' ner ves .
7. The lawyer gave some wise coun sel to his clients.
8. Many people in England mow the la wn of their houses on Sunday morning.
9. The farmer sells the eggs that his he ns lays.
10. Sudden noises at night sca re me a lot.

PVLT for B2 Participants: 5,000- and UWL Word Levels (Test Time 1)

CODE :

11. France was proc laimed a republic in the 18th century.
12. Many people are inj ured in road accidents every year.
13. Suddenly, he was thru into the dark room.
14. He perc eives a light at the end of the tunnel.
15. Children are not independent. They are att ached to their parents.
16. She showed off her sle figure in a long narrow dress.
17. She has been changing partners often because she cannot have a sta ble relationship with one person.
18. You must wear a bathing suit on a public beach. You're not allowed to be na ked.

The 5000-word level

1. Soldiers usually swear an oa of loyalty to their country.
2. The voter placed the ball ot in the box.
3. They keep their valuables in a vau at the bank.
4. A bird perched at the window led .
5. The kitten is playing with a ball of ya .
6. The thieves have forced an ent into the building.
7. The small hill was really a burial mou ntain .
8. We decided to celebrate new year's e ve together.
9. The solider was asked to choose between infantry and cav .
10. This is a complex problem which is difficult to compr omise .
11. The angry crowd sho ut at the prisoner as he was leaving the court.
12. Don't pay attention to this rude remark. Just ign ore it.
13. The management held a secret meeting. The issues discussed were not diss to the workers.
14. We could hear the sergeant bel commands to the troops.
15. The boss got angry with the secretary and it took a lot of tact to soo him.
16. We do not have adequate information to make a decision.
17. She is not a child, but a mat ure woman. She can make her own decisions.
18. The prisoner was put in soli tary confinement.

The University Word List Level

1. There has been a recent tr among prosperous families towards a smaller number of childre
2. The ar ea of his office is 25 square meters.
3. Phil osoph examines the meaning of life.
4. According to the communist doc trine, workers should rule the world.
5. Spending many years together deepened their inti macy .
6. He usually read the sport sec tion of the newspaper first.
7. Because of the doctors' strike the cli nique is closed today.

PVLT for B2 Participants: 10,000-Word Levels (Test Time 1)

CODE :

8. There are several misprints on each page of this te xt .
9. The suspect had both opportunity and mot ivation to commit the murder.
10. They insp ect all products before sending them out to stores.
11. A considerable amount of evidence was accum ulated during the investigation.
12. The victim's shirt was satu with blood.
13. He is irresponsible. You cannot re ly on him for help.
14. It's impossible to eva luate these results without knowing about the research methods that were used.
15. He finally att empts a position of power in the company.
16. The story tells us about a crime and subs punishment.
17. In a hom ogenous class all students are of a similar proficiency.
18. The urge to survive is inh erent in all creatures.

The 10000-word level

1. The baby is wet. Her dia pper needs changing.
2. The prisoner was released on par .
3. Second year University students in the US are called soph .
4. Her favorite flowers were or chidees .
5. The insect causes damage to plants by its toxic sec retions .
6. The evac uation of the building saved many lives.
7. For many people, wealth is a prospect of unimaginable felic .
8. She found herself in a pred without any hope for a solution.
9. The deac helped with the care of the poor of the parish.
10. The hurricane whi along the coast.
11. Some coal was still smol among the ashes.
12. The dead bodies were muti lated beyond recognition.
13. She was sitting on a balcony and bas in the sun.
14. For years waves of invaders pill towns along the coast.
15. The rescue attempt could not proceed quickly. It was imp by bad weather.
16. I wouldn't hire him. He is unmotivated and indo .
17. Computers have made typewriters old-fashioned and obs .
18. Watch out for his wil tricks.

PVLT for B2 Participants: 2000- and 3000-Word Levels (Test Time 2)

Instruction: Complete the underlined words. The example has been done for you.

He was riding a bicycle.

The 2000-word level

1. It is the decision that counts, not the thought.
2. Plants receive water from the soil through their roots.
3. The nurse, was helping the doctor in the operation room.
4. Since he is unskilled, he earns low wages.
5. This year long skirts, are fashionable again.
6. Laws are based upon the principle of justice.
7. He is walking on the tip of his toes.
8. The mechanic had to replace the motor of the car.
9. There is a copy of the original report in the file.
10. They had to climb a steep mountain to reach the cabin.
11. The doctor examinate the patient thoroughly.
12. The house was surrounded by a big garden.
13. The railway connects London with its suburbs.
14. She wandered aimlessly in the street.
15. The organisers listed the number of participants to fifty.
16. This work is not up to your usual standard.
17. They sat down to eat even though they were not hungry.
18. You must have been very brave to participate in such a dangerous operation.

The 3000-word level

1. I live in a small apartment on the second floor.
2. The probability of failing the test scared him.
3. Before writing the final version, the student wrote several drafts.
4. It was a cold day. There was a ch in the air.
5. The cart is pulled by an o
6. Anthropologists study the structures of ancient societies.
7. After two years in the Army, he received the rank of lieutenant.
8. The statue is made of marble.
9. Some aristocrats believed that blue blood flowed through their veins.
10. The secretary assists the boss in organizing the course.
11. His beard was too long. He decided to tr it.
12. People were whirling round on the dance floor.
13. He was on his knees, pleading for mercy.
14. You'll sn that branch if you bend it too far.
15. I won't tell anybody. My lips are sea.
16. Crying is a normal response to pain.
17. The Emperor of China was the supreme ruler of his country.
18. You must be aware that very few jobs are available.

PVLT for B2 Participants: 5000- and UWL Word Levels (Test Time 2)

1. Some people find it difficult to become independent. Instead they prefer to be tied to their mother's ap ron strings.
2. After finishing his degree, he entered upon a new phase in his career.
3. The workmen cleaned up the mess before they left.
4. On Sunday, in his last sermon in Church, the priest spoke against child abuse.
5. I saw them sitting on stool at the bar drinking beer.
6. Her favorite musical instrument was a trumpet.
7. The building is heated by a modern heating apartment.
8. He received many compliments on his dancing skill.
9. People manage to buy houses by raising a mor from a bank.
10. At the bottom of a blackboard there is a le for chalk.
11. After falling off his bicycle, the boy was covered with bruises.
12. The child was holding a doll in her arms and hugging it.
13. We'll have to be inventive and develop a scheme for earning more money.
14. The picture looks nice; the colours blend really well.
15. Nuts and vegetables are considered who food.
16. The garden was full of fragile flowers.
17. Many people feel depressed and gl about the future of the mankind.
18. He is so depressed that he is cont suicide.

The University Word List Level

1. I've had my eyes tested and the optician says my vision is good.
2. The anomalia of his position is that he is the chairman of the committee, but isn't allowed to vote.
3. In their geography class, the children are doing a special pro on North America.
4. In a free country, people can apply for any job. They should not be discriminated against on the basis of colour, age, or sex.
5. A true democracy should ensure equal rights and opportunities for all citizens.
6. The drug was introduced after medical results indisputably proved its effectiveness.
7. These courses should be taken in sequences, not simultaneously.
8. Despite his physical condition, his intellect was unaffected.
9. Governments often cut budgets in times of financial crises.
10. The job offer sounded interesting at first. But when he realised what it would involve, his excitement subs gradually.
11. Research indices that men find it easier to give up smoking than women.
12. In a lecture, most of the talking is done by the lecturer. In a seminar, students are expected to participate in the discussion.
13. The airport is far away. If you want to ensure that you catch your plane, you have to leave early.
14. It's difficult to assert a person's true knowledge by one or two tests.
15. The new manager's job was to restore the company to its former profitability.
16. Even though the student didn't do well on the midterm exam, he got the highest mark on the final.
17. His decision to leave home was not well thought out. It was not based on rat considerations.
18. The challenging job required a young, successful and dynamic candidate.

PVLT for B2 Participants: 10000-Word Levels (Test Time 2)

The 10000-word level

1. The new vic ount was appointed by the bishop.
2. If your lips are sore, try lip sal _____, not medicine.
3. Much to his chag rin, he was not offered the job.
4. The actors exchanged ban alities with reporters.
5. She wanted to marry nobility: a duke, a baron, or at least a vis_____.
6. The floor in the ballroom was a mos_____ of pastel colours.
7. She has contributed a lot of money to various charities. She is known for her generosity and bene_____.
8. This is an unusual singer with a range of three oct_____.
9. A thro_____ controls the flow of gas into an engine.
10. Anyone found loo_____ bombed houses and shops will be severely punished.
11. The crowd soon disp_____ when the police arrived.
12. The wounded man squi_____ on the floor in agony.
13. The dog crin_____ when it saw the snake.
14. He imme_____ himself in a hot bubbly bath forgetting all his troubles for a moment.
15. The approaching storm stam_____ the cattle into running wildly.
16. The problem is beginning to assume mam_____ proportions.
17. His vind_____ behaviour towards the thief was understandable.
18. He was arrested for illi_____ trading in drugs.

Appendix 11 Samples of B2 Participants' Response to the LFP at 0 and 3 Months

(Chapter 5)

LFP for B2 Participants (Test Time 1)

MONT-T1-50

Write about the following topic:

The threat of nuclear weapons maintains world peace. Nuclear power provides cheap and clean energy.

The benefits of nuclear technology far outweigh the disadvantages.

To what extent do you agree or disagree?

Give reasons for your answer and include any relevant examples from your knowledge or experience.

Write at least 250 words.

I agree that nuclear power provides cheap and clean energy. Nevertheless, I disagree that the threat of nuclear technology such as atomic bombs maintains the world in "peace". Indeed, despite the nuclear weapons threat on countries which would want to attempt on the global status quo, we can observe that there are still areas on Earth threatened by war and chaos. Moreover, countries having the nuclear power are not fundamentally led by reliable governments who we can trust in. Furthermore, nuclear incidents such as in Tchernobyl emphasizes on the negative sides of the nuclear technology. In my opinion, nuclear power could be a great thing if we knew how to avoid such catastrophes, and if we don't use its threat as a Damocles's sword on any global leader who would threaten world peace. As far as I know, if a leader wants to enter in war with another country, he will do it, no matter the outcome will be.

LFP for B2 Participants (Test Time 2)

Write about the following topic:

> *The first car appeared on British roads in 1888. By the year 2020 there may be as many as 35 million vehicles on British roads.*
>
> *Alternative forms of transport should be encouraged and international laws introduced to control car ownership and use.*
>
> *To what extent do you agree or disagree?*

Give reasons for your answer and include any relevant examples from your knowledge or experience.

Write at least 250 words.

I agree with the fact that nowadays there is a kind of car invasion, and ~~years~~ years by years there are more and more cars on the Roads as there are more and more people. However it is also a problem of industrialisation, production and excessive consumption.

I personally don't have my driving licence yet, so I didn't have a car. However I am using alternative transport as bus or tramway or even ~~I~~ I just walk when I don't need to go too far.
I think that some people used they car ~~[illegible]~~ by "habits" People are use to go shopping or go to work with their cars so they use it all the times while sometimes they could just need to walk for ~~at least~~ 15 minutes.

The idea of encouraging and even making a laws about car use should be a good thing as it could help to Reduce environemental issues.

If people Reduce their use of cars when it's not necesary and privileges alternative ways it will be a better thing for environement (air) and also for the circulation in cities, it will Reduce the invasion of Roads.

Moreover in some ~~places~~ people are far away from accesible shops etc... so it will be a good think to invest on bus lines extention ~~[illegible]~~ for the people to get access to it and leave their cars when it's not necessary.
By living in Montpellier for 3 years now, I think that it is well organised. You don't need to use your car to go to one point of the city to another and that is a really good thing especially for carpark.
The solution need to be introduced before it is ~~too~~ late.

References

Bachman, L. (1990). *Fundamental Considerations in Language Testing.* Oxford University Press. https://doi.org/10.2307/329499

Bachman, L., & Palmer, A. (1996). *Language testing in practice.* Oxford University Press.

Barrow, J., Nakanishi, Y., & Nishino, H. (1999). Assessing Japanese college students' vocabulary knowledge with a self-checking familiarity survey. *System, 27*(2), 223-247.

Boersma, P., & Weenink, D. (2005). *PRAAT: doing phonetics by computer*. https://www.fon.hum.uva.nl/praat/ Chapelle, C. (2006). L2 vocabulary acquisition theory: The role of inference, dependability and generalizability in assessment. In M. Chalhoub-Deville, C. Chapelle & P. Duff, *Inference and generalizability in applied linguistics: multiple perspectives* (pp. 47–64). John Benjamins.

Clenton, J. (2010). Investigating the construct of productive vocabulary knowledge with Lex30. *(Unpublished Doctoral Dissertation). University of Swansea.*

Clenton, J. (2015). Testing the Revised Hierarchical Model: Evidence from word associations. *Bilingualism: Language and Cognition, 18*(1), 118-125. https://doi.org/10.1017/S136672891400008X

Clenton, J., De Jong, N., Clingwall, D., & Fraser, S. (2021). Investigating the extent to which vocabulary knowledge and skills can predict aspects of fluency for a small group of pre-intermediate Japanese L1 users of English (L2). In J. Clenton, & P. Booth, *Vocabulary and the Four Skills Pedagogy, Practice, and Implications for Teaching Vocabulary* (pp. 126-145). Routledge.

Crossley, S., Salsbury, T., & McNamara, D. (2009). Measuring l2 lexical growth using hypernymic relationships. *Language Learning, 59*(2), 307-334. https://doi.org/10.1111/j.1467-9922.2009.00508.x

Dang, T., Coxhead, A., & Webb, S. (2017). The Academic Spoken Word List. *Language Learning, 67*(4), 959-997. https://doi.org/10.1111/lang.12253

Dang, T., Webb, S., & Coxhead, A. (2020). Evaluating lists of high-frequency words: Teachers' and learners' perspectives. *Language Teaching Research*, 1-25. https://doi.org/10.1177/1362168820911189

De Jong, N. (2016). Fluency in second language assessment. In D. Tsagari, & J. Banerjee, *Handbook of Second Language Assessment* (pp. 203-218). Mouton de Gruyter.

De Jong, N., & Mora , J. (2017). Does having good articulatory skills lead to more fluent speech in first and second languages? *Studies in Second Language Acquisition, 41*(1), 227-239. https://doi.org/10.1017/S0272263117000389

De Jong, N., Groenhout, R., Schoonen, R., & Hulstijn, J. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics, 26*(2), 223-243.

De Jong, N., Steinel, M., Florijn, A., Schoonen, R., & Hulstijn, J. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34*(1), 5-34. https://doi.org/10.1017/S0272263111000489

Elgort, I. (2018). Technology-Mediated Second Language Vocabulary Development: A Review of Trends in Research Methodology. *CALICO Journal, 35*(1), 1-29. https://doi.org/10.1558/cj.34554

Ellis, N. C. (2002). Frequency effects in language processing: A Review with Implications for Theories of Implicit and Explicit Language Acquisition. *SSLA, 24*(2), 143-188. https://doi.org/10.1017/s0272263102002024

Elmetaher, H. (2021). Investigating Productive Vocabulary Knowledge Development: A Task-Based Approach. *Studies in European and American Cultures Journal,* 28, 1-25.

ETS. (2005). *TOEFL iBT.* ETS. https://www.ets.org/toefl/test-takers/

ETS. (2019). *TOEIC Bridge*. https://www.ets.org/s/toeic/pdf/toeic_bridge_cefr_flyer.pdf

Fitzpatrick, T. (2007). Productive vocabulary tests and the search for concurrent validity. In H. Daller, J. Milton, & J. Treffers-Daller , *Modelling and assessing vocabulary knowledge* (pp. 116-132). Cambridge University Press.

Fitzpatrick, T., & Clenton, J. (2010). The challenge of validation: Assessing the performance of a test of productive vocabulary. *Language Testing, 27*(4), 537-554. https://doi.org/10.1177/0265532209354771

Fitzpatrick, T., & Clenton, J. (2017). Making Sense of Learner Performance on Tests of Productive Vocabulary knowledge. *TESOL Quarterly, 51*(4), 844-867. https://doi.org/10.1002/tesq.356

Fitzpatrick, T., & Meara, P. (2004). Exploring the Validity of a Test of Productive Vocabulary. *Vigo International Journal of Applied Linguistics, 1*, 55-74.

Fountain, R., & Nation, I. (2000). A vocabulary-based graded dictation test. *RELC Journal, 3*(2), 29-44. https://doi.org/10.1177/003368820003100202

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

Gu, P. (2019). Strategies for Learning Vocabulary. In S. Webb, *The Routledge Handbook of Vocabulary Studies.* Routledge.

Heatley, A., & Nation, I. (1998). *VocabProfile and Range.* School of Linguistics and Applied Language Studies, Victoria University of Wellington.

Henriksen, B. (1999). Three Dimensions of Vocabulary Developmen. *Studies in Second Acquisition, 21*(2), 303-317. https://doi.org/10.1017/s0272263199002089

Higgs, T., & Clifford, R. (1982). The push toward communication. In T. Higgs, *Curriculum, competence and the foreign language teacher* (pp. 243-265). National Textbook Company.

Horst, M., & Collins, L. (2006). From faible to strong: How does their vocabulary grow? *The Canadian Modern Language Review/La Revue canadienne des langues vivantes, 63*(1), 83-106.

Housen, A., Bulte, B., Pierrard, M., & Daele, S. (2008). Investigating lexical proficiency development over time - the case of Dutch-speaking learners of French in Brussels. *French Language Studies, 18*, 1-22. https://doi.org/10.1017/S0959269508003451

IELTS. (2019a). *Sample test questions*. *IELTS*. https://www.ielts.org/for-test-takers/sample-test-questions

IELTS. (2019b). *IELTS grows to 3.5 million a year*. *IELTS.org*. https://www.ielts.org/news/2019/ielts-grows-to-three-and-a-half-million-a-year

IELTS. (2019c). *Test taker performance 2019*. *IELTS.org*. https://www.ielts.org/for-researchers/test-statistics/test-taker-performance

IELTS. (2021a). *IELTS TASK 2 Writing band descriptors (public version)*. *British Council.* https://takeielts.britishcouncil.org/sites/default/files/ielts_task_2_writing_band_descriptors.pdf

IELTS. (2021b). *British Council*. SPEAKING: Band Descriptors (public version): https://www.ielts.org/-/media/pdfs/speaking-band-descriptors.ashx?la=en

Ishikawa, S., Uemura, T., Kaneda, M., Shmizu, S., Sugimori, N., Tono, Y., Murata, M. (2003). *JACET 8000: JACET List of 8000 Basic Words.* JACET.

Jiang, N. (2000). Lexical Representation and Development in a Second Language. *Applied Linguistics, 21*(1), 47-77. https://doi.org/10.1093/applin/21.1.47

Jiang, N. (2002). Form-meaning mapping in vocabulary acquisition in a second language. *Studies in Second Language Acquisition, 24*(4), 617-637. https://doi.org/10.1017/S0272263102004047

Jiang, N. (2004). Semantic Transfer and Its Implications for Vocabulary Teaching in a Second Language. *The Modern Language Journal, 4*, 416-432.

Johnson , M., Acevedo , A., & Mercado, L. (2016). Vocabulary Knowledge and Vocabulary Use in Second Language Writing. *TESOL, 7*(3), 700-715. https://doi.org/10.1002/tesj.238

Kiss, G., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In A. Aitken, R. Bailey, & N. Hamilton-Smith, *The Computer and Literary Studies* (pp. 153-165). Edinburgh University Press.

Laufer, B. (1994). The Lexical Profile of Second Language Writing: Does It Change Over Time? *RELC Journal, 25*(2). https://doi.org/10.1177/003368829402500202

Laufer, B. (1998). The Development of Passive and Active Vocabulary in a Second Language: Same or Different? *Applied Linguistics, 19*(2), 25S-27I. https://doi.org/10.1093/applin/19.2.255

Laufer, B., & Goldstein, Z. (2004). Testing Vocabulary Knowledge: Size, Strength, and Computer Adaptiveness. *Language Learning, 54*(3), 399-436. https://doi.org/10.1111/j.0023-8333.2004.00260.x

Laufer, B., & McLean, S. (2016). Loanwords and Vocabulary Size Test Scores: A Case of Different Estimates for Different L1 Learners. *Language Assessment Quarterly, 13*(3), 202-217.

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

Laufer, B., & Nation, P. (1999). A vocabulary -size test of controlled productive ability. *Applied Linguistics, 16*(1), 33-51. https://doi.org/10.1177/026553229901600103

Laufer, B., & Paribakht, T. S. (1998). The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning, 48*(3), 365-391. https://doi.org/10.1111/0023-8333.00046

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40*, 387-417.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk.* Lawrence Erlbaum Associates.

McLean, S., & Kramer, B. (2015). The Creation of a New Vocabulary Levels Test. *Shiken, 19*(2), 1-11.

McLean, S., Hogg, N., & Kramer, B. (2014). Estimations of Japanese university learners' english vocabulary sizes using the vocabulary size test. *Vocabulary Learning and Instruction, 3*(2), 47-55. https://doi.org/10.7820/vli.v03.2.mclean.et.al

Meara, P. (1990). A note on passive vocabulary. *Second Language Research, 6*(2), 150-154. https://doi.org/10.1177/026765839000600204

Meara, P. (1997). Towards a New Approach to Modelling Vocabulary Acquisition. In N. Schmitt, & M. McCarthy, *Vocabulary: Description, Acquisition and Pedagogy* (pp. 109-121). Cambridge University Press.

Meara, P., & Alcoy, O. (2010). Words as Species: An Alternative Approach to Estimating Productive Vocabulary Size. *Reading in a Foreign Language, 28*(1), 222-236. https://doi.org/10.1016/S0346-251X(99)00058-5

Meara, P., & Fitzpatrick, T. (2000). Lex30: An improved method of assessing productive vocabulary in an L2. *System, 28*(1), 19-30. https://doi.org/10.1016/S0346-251X(99)00058-5

Meara, P., & Milton, J. (2003). *X_Lex, The Swansea Levels Test.* Express.

Melka, F. (1997). Receptive versus productive aspects of vocabulary. In Schmitt, & McCarthy, *Vocabulary: Description, acquisition, and pedagogy* (pp. 84-102). Cambridge University Press.

Milton, J. (2009). *Measuring Second Language Vocabulary Acquisition.* Multilingual Matters.

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist, & B. Laufer, *L2 vocabulary acquisition,*

knowledge and use. New perspectives on assessment and corpus analysis (Vol. Eurosla Monographs Series 2, pp. 57-78). Eurosla.

Miralpeix, I. (2019). *L1 and L2 Vocabulary Size and Growth.* Routledge.

Nation, I., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9-13.

Nation, P. (1984). *Vocabulary lists: Words, affixes, and stems.* Victoria University of Wellington.

Nation, P. (2001). *Learning vocabulary in another language.* Cambridge University.

Nation, P. (2017). *The BNC/COCA Level 6 word family lists (Version 1.0.0) [Data file]*. http://www.victoria.ac.nz/lals/staff/ paul-nation.aspx

Nation, P., & Webb, S. (2011). *Researching and Analyzing Vocabulary.* Heinle.

Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics, 25*, 26-45.

Palmberg, R. (1987). Patterns of Vocabulary Development in Foreign-Language Learners. *Studies in Second Language Acquisition, 9*(2), 201-219. https://doi.org/10.1017/S0272263100000474

Paribakht, T., & Wesche, M. (1993). Reading comprehension and second language development in a comprehension-based ESL program. *TESL Canada Journal, 11*(1), 9-29. https://doi.org/10.18806/tesl.v11i1.623

Paribakht, T., & Wesche, M. (1996). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady , & T. Huckin , *Second Language Vocabulary Acquisition: A Rationale for Pedagogy* (pp. 174-200). Cambridge University Press. https://doi.org/10.1017/cbo9781139524643.013

Pellicer-Sánchez, A. (2018). Examining second language vocabulary growth: Replications of Schmitt (1998) and Webb & Chang (2012). *Cambridge University Press, 52*(4), 512-523. https://doi.org/10.1017/s026144481800037x

Read, J. (2000). *Assessing vocabulary knowledge and use.* Cambridge University Press.

Read, J. (2012). Piloting vocabulary tests. In G. Fulcher, & F. Davidson, *The Routledge Handbook of Language Testing* (pp. 307-320). Routledge.

Schmitt, D. (2010). *Researching Vocabulary: A Vocabulary Research Manual.* Palgrave Press.

Schmitt, N. (2019). Understanding vocabulary acquisition, instruction, and assessment: A research. *Language Teaching, 52*, 261-274.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002a). *E-prime reference guide.* Psychology Software Tools.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002b). *E-prime user's guide.* Psychology Software Tools.

Segalowitz, N. (2010). *Cognitive bases of second language fluency.* Routledge.

Shillaw, J. (1995). Using a word list as a focus for vocabulary learning. *The Language Teacher, 19*(2), 58-59.

Stæhr, L. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition, 31*(4), 577-607. https://doi.org/10.1017/s0272263109990039

Union, D. L. (2004). *Corpus of spoken Dutch*. http://lands.let.ru.nl/cgn/ehome.htm

Walters, J. (2012). Aspects of validity of a test of productive vocabulary: Lex30. *Language Assessment Quarterly, 9*(2), 172-185. https://doi.org/10.1080/15434303.2011.625579

Webb, S. (2005). Receptive and Productive Vocabulary Learning: The Effects of Reading and Writing on Word Knowledge. *Studies in Second Language Acquisition, 27*(1), 33-52. https://doi.org/10.1017/s0272263105050023

Webb, S. (2007). The Effects of Synonymy on Second-Language Vocabulary Learning. *Reading in a Foreign Language, 19*(2), 120–136.

Webb, S. (2009). The Effects of Receptive and Productive Learning of Word Pairs on Vocabulary Knowledge. *RELC Journal, 4*(3), 360-376. https://doi.org/10.1177/0033688209343854