Exploring the Construct Validity of Tests Used to Assess L2 Productive Vocabulary Knowledge

Amanda Edmonds, Jon Clenton, Hosam Elmetaher

Exploring the Construct Validity of Tests Used to Assess L2 Productive Vocabulary Knowledge

Amanda Edmonds,[a] Jon Clenton,[b] Hosam Elmetaher[c]

[a]Université Côte d'Azur

[b]Hiroshima University

[c]Nanzan University

This work was supported by the JSPS [project number 19K00881]

Edmonds, A., Clenton, J., & Elmetaher, H. (2022). Exploring the construct validity of tests used to assess L2 productive vocabulary knowledge. *System*. Advance online publication. https://doi.org/10.1016/j.system.2022.102855

Exploring the Construct Validity of Tests Used to Assess L2 Productive Vocabulary Knowledge

Abstract

Vocabulary knowledge in a second language (L2) is thought to be a complex construct and, accordingly, there exist numerous ways to evaluate a L2 learner's vocabulary knowledge. These assessments are generally billed as tests of vocabulary size (i.e., number of words known) or depth (i.e., how well words are known) of either receptive or productive vocabulary knowledge. However, inconsistencies persist in how existing assessments are characterized, leading to sometimes contradictory claims over what these tests are measuring. This state of affairs leads to concerns with the construct validity of the tests in question. In the current study, we contribute to these ongoing discussions with a focus on tests that target productive vocabulary knowledge in L2 English. Four vocabulary tests (three productive, one receptive) were administered to a group of Francophone learners of English, and results were analyzed using an exploratory factor analysis. This analytic approach led to the identification of two underlying constructs, which we labeled receptive vocabulary knowledge and productive vocabulary knowledge, respectively. These results highlight the importance of the crucial distinction between receptive and productive knowledge in the conceptualization of the overall construct of vocabulary knowledge.

Keywords: vocabulary, productive, receptive, exploratory factor analysis

1. Introduction

Within the field of second language acquisition (SLA), vocabulary knowledge has received sustained theoretical and empirical attention over the past three decades (e.g., Webb, 2020), with second language (L2) vocabulary researchers consistently highlighting the complex nature of vocabulary knowledge. In particular, two main conceptual distinctions cut

across and structure relevant research. First, vocabulary knowledge is commonly conceived of either in terms of the number of words that a learner knows (i.e., their vocabulary size), or in terms of how well words are known (i.e., vocabulary depth). A second important distinction contrasts receptive (or passive) knowledge of words with productive (or active) knowledge. Despite continuing debates over their operationalization, the concepts of vocabulary size versus depth and receptive versus productive knowledge have proven useful in L2 vocabulary research, particularly in assessments of vocabulary knowledge. Indeed, taking account of these two concepts suggests that vocabulary knowledge may rely on four separate constructs – receptive vocabulary size, receptive vocabulary depth, productive vocabulary size, and productive vocabulary depth –, each of which may presumably be assessed using specific tests. However, the potential for a clean mapping of assessment measures onto these four potential constructs is belied by the characterizations and actual uses of many tests. Two examples are illustrative: (i) although there is agreement that Lex30 (Meara & Fitzpatrick, 2000) elicits productive vocabulary knowledge, disagreements arise as to whether it provides a measure of depth of vocabulary knowledge (see Read, 2004, p. 220) or vocabulary size (i.e., Williams, Segalowitz & Leclair, 2014); and, (ii) although the Productive Vocabulary Levels Test (PVLT, Laufer & Nation, 1999) was designed to reveal the size of an individual's productive vocabulary, given that participants must complete words presented in carrier sentences, Webb (2005, p. 82) suggests that it might "actually test receptive knowledge." These examples highlight ongoing discussions about the construct validity of vocabulary tests, namely discussions of whether they measure what they purport to. Accordingly, in the current study, we contribute to these discussions with a project whose point of departure was tests that target productive vocabulary knowledge in L2 English. This focus is justified by the relative paucity of research on the assessment of productive (versus receptive) vocabulary knowledge. We thus administered four vocabulary tests (three productive and one receptive)

to a group of 100 Francophone learners of L2 English. While previous research (see, e.g., Clenton, 2010; Fitzpatrick, 2007; Fitzpatrick & Clenton, 2017; Walters, 2012) has reflected on the construct validity of multiple vocabulary tests using correlational analyses, we opt for a different analytic approach. More specifically, we analyzed our results using exploratory factor analysis, an approach that allows the researcher to explore complex correlational patterns in parallel. This allowed us to uncover two constructs that underpin the performances on the four tests, thus offering novel insights into the ongoing debates over the construct validity of different L2 vocabulary tests.

2. Background

2.1 *Conceptualizing Vocabulary Knowledge in an L2*

Two distinctions are central in conceptualizing vocabulary knowledge in a L2: vocabulary size/depth and receptive/productive knowledge. Vocabulary size refers to the number of words[1] that an individual knows. In most research, knowing a word has been synonymous with demonstrating receptive knowledge about its form-meaning mapping, and it has been assumed that more frequent words are learned earlier and better than less frequent ones (see Milton, 2007, for discussion). As a result, most assessments of vocabulary size target a sample of words from various frequency bands and extrapolate size scores from performance on the sample. Vocabulary depth, on the other hand, is a more unruly concept to define, with numerous approaches having been proposed (see Read, 2004; Schmitt, 2014; Yanagisawa & Webb, 2020). In simple terms, vocabulary depth refers to how well one knows a word; however, the operationalization of this concept often proves challenging. One common way of conceiving of vocabulary depth has come to be known as the components (or dimensions) approach. Although this approach to vocabulary depth can take many forms, they

[11] In this paper we refer to the generic term 'word'. While we acknowledge the complexity and appropriacy of different lexical units (e.g., Webb, 2021), we use word types as our measurement unit.

all agree that vocabulary knowledge covers a variety of components. Nation's (2013) list of nine components of vocabulary knowledge is a well-known example of such an approach:

(1) Component 1: spoken form

Component 2: written form

Component 3: word parts

Component 4: form and meaning

Component 5: concept and referents

Component 6: associations

Component 7: grammatical functions

Component 8: collocations

Component 9: constraints on use

In a components approach, knowledge with respect to more dimensions is considered to reflect greater depth of knowledge. A second way to conceive of vocabulary depth is what Read (2004) referred to as network knowledge. In this case, greater vocabulary depth is equated with words boasting richer associative networks. More elaborated connective networks are interpreted as reflecting greater lexical organization (Meara & Wolter, 2004). Considering vocabulary size and depth together, numerous studies have reported evidence of a strong positive relationship between the two (see Schmitt, 2014, for an overview). This is not particularly surprising, given that "size by definition is the number of lexical items known to some criterion level of mastery. But the criterion will always be some measure of depth" (Schmitt, 2014, p. 942). Schmitt goes further, stating that "there can be no clear distinction between size and depth." Following this reasoning, it is unsurprising that inconsistencies may exist concerning whether certain tests tap vocabulary size or depth.

The second central conceptual distinction in L2 vocabulary research concerns receptive versus productive knowledge. Generally speaking, receptive abilities correspond to

the recognition and understanding of words, whereas productive abilities include the capacity to recall or produce a word. If this distinction is generally accepted by vocabulary researchers, there continue to be discussions about the underlying nature of the distinction. On the one hand, Meara (1997) discusses the possibility of receptive and productive abilities being qualitatively different. This qualitative difference depends, according to Meara, on the associative links attached to a word in the mental lexicon. Words for which a speaker has receptive knowledge have fewer links and need external stimuli to be activated. Words for which one has productive knowledge can be activated within the mental lexicon (independent of external stimuli), by dint of their richer associative links. Other researchers conceive of receptive-productive (or passive-active) knowledge as a continuum, and this continuum has formed the foundation for many conceptualizations of vocabulary knowledge in a L2 (e.g., Henriksen, 1999; Laufer & Goldestein, 2004; Melka, 1997; Palmberg 1987; Schmitt, 2010). This view suggests that language users need to develop their receptive vocabulary knowledge to the extent that it becomes productive, and an actively used component of the learner lexicon. Although the exact threshold at which receptive knowledge becomes productive has not as yet been identified (Laufer & Goldstein 2004; Schmitt 2010), Schmitt (2019, p. 264) suggests that "learning most words to receptive mastery is relatively easy; it is enhancing that knowledge to productive mastery which is the real challenge." Empirical results concerning the acquisition of receptive versus productive vocabulary knowledge indicate that "receptive vocabulary knowledge develops before and at a faster rate than productive vocabulary, is larger than productive vocabulary and, importantly, is easier and more straightforward to measure or quantify than its productive counterpart" (Williams et al., 2014, p. 24). Concerning Williams et al.'s observation that receptive knowledge "is easier and more straightforward to measure or quantify than its productive counterpart", there indeed currently

exist more assessments of receptive knowledge than there do of productive knowledge (see Miralpeix, 2020, p. 192).

2.2 *Assessing Productive Vocabulary Knowledge*

A small number of tools have been developed in order to assess productive vocabulary knowledge in a L2. For the current project, we focused on three such measures: (i) Lex30 (Meara & Fitzpatrick, 2000) based on word association, using single word cues, (ii) G_Lex (Fitzpatrick & Clenton, 2017) based on word association, using sentence cues, and (iii) the PVLT (Laufer & Nation 1999), a sentence completion test. Example items for each of the tests, along with instructions, are provided in Table 1.

Table 1

Instructions and Example Items for Three Tests of Productive Vocabulary Knowledge

Measures	Instructions	Example items
Lex30	Write down the first four (English) words you think of when you read each word in the list	1. attack 2. board 3. close
G_Lex	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 _____.
PVLT	Complete the underlined words.	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_____.

On the basis of the example items, it is clear that each of these three measures takes a rather different approach to eliciting productive vocabulary knowledge. Beginning with Lex30, Meara and Fitzpatrick (2000) opted for a word association format. Respondents are asked to provide up to four words in response to 30 single-word stimuli (see Figure 1A in the Appendix for an example of a completed Lex30 test). Completed Lex30 papers are typed up, lemmatized, and then compared with a corpus to determine a Lex30 score. All function words, proper nouns, numbers, and those words that fall within the first 1,000 frequency band do not score. Thus, a Lex30 score consists of a count of all but the highly frequent (i.e., non-1000) responses. Lex30 scores have been expressed as either a simple count of the infrequent items or the percentage of infrequent items. To date, this test has been used in numerous empirical studies (e.g., Clenton, 2015; Clenton, de Jong, Clingwall & Fraser, 2020; Elmetaher, 2021; Gilyuk, Edmonds & Sneed German, 2021; Uchihara, Eguchi, Clenton, Kyle & Saito, 2021) and has been subjected to several validation attempts (see Fitzpatrick & Clenton, 2010; Fitzpatrick & Meara, 2004; Walters, 2012).

G_Lex was devised by Fitzpatrick and Clenton (2017) as a point of comparison to investigate the construct underlying Lex30. G_Lex is a sentence completion test in which participants are asked to provide up to five words to complete each of 24 sentence gaps (an example is provided in the Appendix, Figure 1B). Like Lex30, no specific responses are targeted, and 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. The main difference between G_Lex and Lex30 is that the former encourages test takers to consider context (both in terms of part of speech and appropriate meaning) in providing their responses. Unlike the

other tests used in this study, G_Lex has not been as extensively tested, and thus our study serves to provide initial evidence of the construct validity of this relatively new test.

The PVLT (Laufer & Nation, 1999) – was designed to provide an estimate of productive vocabulary size using a word completion format (see Appendix, Figure 1C, for an example). On the PVLT, 18 test sentences for each of five frequency levels (2k, 3k, 5k, UWL [the University Word List], and 10k) are presented for a total of 90 items, and in each sentence, participants are required to provide a target word (belonging to the frequency band in question). The first few letters of each target item are provided to restrict responses to a specific target word. A point is awarded for each correctly provided target item. The PVLT is a widely used measure of productive vocabulary knowledge (e.g., de Jong et al., 2012; Eguchi, Suzuki, & Suzuki, 2021), and has been the subject of validation tests (Laufer, 1998; Laufer & Nation, 1999).

Each of these tests aims to assess productive vocabulary knowledge. However, whereas Laufer and Nation (1999) clearly state that the PVLT is intended as a test of productive vocabulary size, it is less clear whether the creators of Lex30 and especially G_Lex intended their tests as measures of size, depth, or perhaps as a general measure of productive vocabulary (for Lex30, see Meara & Fitzpatrick, 2000, p. 22 and 27). Looking beyond the size-depth distinction, authors such as Fitzpatrick and Meara (2004), Fitzpatrick (2007), and Clenton (2010) have argued that existing productive vocabulary measures have the potential to mobilize – and, thus, may measure – quite different aspects of vocabulary knowledge. Using Nation's (2013) nine aspects of vocabulary knowledge, these researchers argue that whereas all tests can tap into the productive form-meaning connection (which is generally how vocabulary size is operationalized), they vary considerably in the other types of vocabulary knowledge (i.e., other dimensions of vocabulary depth) that can be demonstrated. Fitzpatrick and Meara (2004, p. 72) go so far as to suggest that the assessments of productive

vocabulary knowledge they explored "do not measure the same things at all; productive vocabulary is a misleadingly simplistic label for an extremely complex construct."

With respect to the three productive vocabulary measures included in the present study, a similar observation can be made. In Table 2, we build on work by Fitzpatrick (2007, p. 129) and Clenton (2010, p. 178) in order to provide an overview of the aspects of vocabulary knowledge that each measure may arguably tap. Although all three tests presumably allow participants to demonstrate productive knowledge of how a word is written and the productive form-meaning connection for a given word, only Lex30 and G_Lex allow learners to showcase productive knowledge of associations and only G_Lex and PVLT potentially tap into receptive collocational knowledge (because of the presence of carrier sentences). The PVLT arguably taps three additional vocabulary knowledge dimensions. Due to the constrained nature of this test and the necessity to be able to understand the carrier sentence in order to provide the targeted response, participants must demonstrate knowledge of grammatical functions and constraints on use. Finally, because the PVLT provides between 1 and 5 letters of the target word, receptive recognition of what the target words look like is also tested (see Read, 2020, pp. 549-550). The profile provided in Table 2 thus suggests that the PVLT may tap three aspects of receptive knowledge, echoing Webb's (2005, p. 82) concern about the construct validity of this test as an assessment of productive vocabulary knowledge. Given that the aim of the current project is to contribute to discussions on the construct validity of productive vocabulary assessments, the fact that PVLT may also be tapping into receptive knowledge led us to include a fourth measure – a receptive vocabulary knowledge test – in the present study. The inclusion of a receptive test importantly allows us to verify whether performance on the PVLT patterns more closely with performance on other productive measures of vocabulary or with the Vocabulary Levels Test (VLT, Nation, 1983; Schmitt, Schmitt & Clapham, 2001), the receptive measure included in this study. The VLT is

a widely used test for assessing receptive vocabulary size (but see Webb, Sasao & Ballance, 2017, for a discussion of the appropriateness of this use; an example from the test is provided in the Appendix, Figure 1D). As is shown in Table 2, the types of knowledge potentially assessed by the VLT overlap with the knowledge types mobilized by the PVLT, insofar as both tests require the recognition of the written form of target words. This overlap, combined with the fact that the PVLT was developed as a productive counterpart to this test, makes the VLT a relevant point of comparison for the PVLT (i.e., these two tests may measure similar aspects of vocabulary knowledge).

Table 2

Dimensions of Vocabulary Knowledge Tapped by Four Vocabulary Tests

				Lex30	G_Lex	PVLT	VLT
Form	Spoken form	R	What does the word sound like?				
		P	How is the word pronounced?				
	Written form	R	What does the word look like?			✓	✓
		P	How is the word written and spelled?	✓	✓	✓	
	Word parts	R	What parts are recognizable in this word?				
		P	What word parts are needed to express the meaning?				
Meaning	Form and meaning	R	What meaning does this word signal?				✓
		P	What word form can be used to express this meaning?	✓	✓	✓	

	Concept and referents	R	What is included in the concept?				
		P	What items can the concept refer to?				
	Associations	R	What other words does this word make us think of?				✓
		P	What other words could we use instead of this one?	✓	✓		
Use	Grammatical functions	R	In what patterns does the word occur?			✓	
		P	In what patterns must we use the word?			✓	
	Collocations	R	What words or types of words occur with this one?		✓	✓	
		P	What words or types of words must we use with this word?				
	Constraints on use	R	Where, when, and how often would we expect to meet this word?				
		P	Where, when, and how often can we use this word?			✓	

Note. R= Receptive, P = Productive

2.3 *Assessing Assessments of L2 Vocabulary Knowledge*

Given the multi-faceted nature of vocabulary knowledge and the growing number of measures proposed to measure that knowledge, researchers have turned their attention to the

important endeavor of exploring the reliability and validity of these different tests (for a given population and for a given context). There are many ways to contribute to this line of research, and in what follows, we briefly review the use of correlations and structural equation modeling (SEM). We end by explaining what exploratory factor analysis, the approach adopted in this study, has to offer.

One approach to determining whether a given test measures the knowledge or skill it was intended to measure is to compare results on the test in question with results obtained from another test that has been shown to measure the construct of interest. This is what is referred to as concurrent validity. According to Langsford et al. (2018, p. 5), "measure agreement is an important check of validity for diverse measures claiming to reflect the same underlying construct." This approach has been frequently adopted in L2 vocabulary measurement studies and has involved exploring the correlations obtained between the results from two vocabulary tests (Bachman, 1990, p. 248). Limiting ourselves to those studies having explored concurrent validity involving at least two of the tests included in the current study, several conclusions can be drawn. First, studies having assessed vocabulary knowledge using the VLT and PVLT report significant positive correlations between the scores (Laufer & Paribakht, 1998; Nemati, 2010). Second, investigations that have administered the PVLT and Lex30 to the same participants have also reported significantly positive correlations (Elmetaher, 2021; Fitzpatrick & Meara, 2004; Walters, 2012). Third, in the two studies to explore the concurrent validity of Lex30 and G_Lex (Elmetaher, 2021; Fitzpatrick & Clenton, 2017), a moderate positive correlation was reported. Taken together, these results provide evidence that the constructs measured by the different tests are highly related. These results are moreover consistent with the possibility that the tests tap the same underlying construct. However, separate bivariate correlations cannot reveal more complex correlational patterns and, as such, may provide an overly simplistic perspective on latent constructs. Other types of

analysis – namely, SEM and factor analysis – are designed to accommodate and identify such complex patterns.

SEM is designed to allow the researcher to test hypothesized relationships among sets of measured and latent variables. In other words, this type of analysis is confirmatory, insofar as researchers propose a model (or a set of models) that they then test against a given dataset (see Schoonen, 2015, for an overview of SEM applied to SLA). In the realm of L2 vocabulary acquisition, a small number of researchers have availed themselves of SEM for various aims: to offer a model of motivated vocabulary learning (Tseng & Schmitt, 2008), to model the impact of vocabulary and grammatical knowledge on reading comprehension (Zhang, 2012), or – and of particular relevance to the present study – "to examine the nature of the overall vocabulary knowledge construct" (González-Fernández & Schmitt, 2020, p. 491). In their study, González-Fernández and Schmitt administered nine vocabulary tests to 144 Spanish-speaking learners of English. These tests included one measure of receptive vocabulary size (the VLT) and eight measures of vocabulary depth (for 20 target words), covering both recall (an aspect of productive knowledge) and recognition (an aspect of receptive knowledge) for four components of vocabulary knowledge: form-meaning connection, derivations, polysemy, and collocations. Spearman correlations between all tests were high and positive. Results from the SEM analysis, however, revealed more complex patterns. Using the results from the eight measures of vocabulary depth, two alternate models of vocabulary knowledge were tested. Model 1 hypothesized that the four components (knowledge of form-meaning connections, derivations, polysemy, and collocations) would be direct contributors to vocabulary knowledge, with the recall and recognition measures for each contributing to the component in question. In Model 2, on the other hand, recall and recognition measures for each component contributed directly to the vocabulary knowledge construct. The authors found that Model 2 was supported by the data, which they suggest indicates "that recognition

vs. recall knowledge was the key distinction in our study" (p. 501). Concretely, this means that, for example, performance on measures of receptive knowledge of different components appeared to behave more similarly than did receptive and productive knowledge for the same component.

SEM and other confirmatory analyses (e.g., confirmatory factor analysis) are appropriate when the researcher wishes to test a predetermined model. Exploratory factor analysis, on the other hand, is a type of analysis that allows scholars to "uncover the latent constructs underlying the variables, in an attempt to better understand the nature of such constructs" (Bandalos & Boehm-Kaufman, 2009, p. 63). Exploratory factor analysis is appropriate in cases "when researchers do not have any particular expectations regarding the number and nature of the underlying factors (i.e., latent variables) that exist in the data" (Loewen & Gonulal, 2015, p. 183). Factor analyses have been widely used to address questions of construct validity, with exploratory factor analysis often used to develop a model that can subsequently be tested using SEM or confirmatory factor analysis. To take one example from SLA, factor analyses have been central in the debate over the construct validity of different measures of implicit and explicit knowledge (see, among others, Ellis, 2005; Ellis & Loewen, 2007; Gutiérrez, 2013). In the context of the present study, we are interested in exploring the construct validity of four measures of L2 vocabulary knowledge. Exploratory factor analysis (as opposed to a confirmatory approach) was deemed most appropriate considering current discussions and debates regarding the nature of vocabulary knowledge and the measurement of that knowledge within the broader context of the L2. In particular, we note that the characterization of assessments of vocabulary knowledge continues to be the subject of debate. Not only are there disagreements in classification according to the two most widely cited dimensions (i.e., vocabulary size versus depth and receptive versus productive knowledge), but researchers such as Fitzpatrick and Meara (2004, p. 72) have explicitly stated

that tests of productive vocabulary knowledge may not measure the same construct, while others (Fitzpatrick & Clenton, 2017) have highlighted that test constructs are to some extent overlapping. While attempts have been made to identify different aspects of vocabulary knowledge potentially tapped by different tests (see Table 2), there is as yet no clear model to test.

3. Method

3.1 *Participants*

In selecting our participant population, we opted for highly proficient learners, given that some of the words targeted by two of the tests (PVLT and VLT) are infrequent and, thus, unlikely to be known by less proficient learners. Data were collected from all students enrolled in the fourth semester of an English degree program at a French university. Overall, these participants were highly proficient in the L2, as they were able to function in an academic context where most classes and classwork were conducted in English. Each semester, students enrolled in this program take obligatory courses on translation, English linguistics, Anglophone history and culture, and English literature, for a total of 12 to 15 hours a week of instruction in English. Our analysis is based on data from the 100 participants who completed all measures. On average, participants were 20.47 years old (*SD* = 2.04, range = 18-28), and 79 were women. Participants reported speaking a wide variety of languages within their families: 63 reported speaking only French, 29 reported speaking French and at least one other language, 7 reported growing up speaking a language other than French (e.g., Berber, Wolof), and one person did not respond to this question. When asked what foreign languages they had studied, six participants responded only English, 38 reported learning English and one other foreign language, and all remaining participants stated that they had learned English in addition to two or more foreign languages.

3.2 *Data Collection*

Participants completed four English vocabulary measures: Lex30, G_Lex, the PVLT, and the VLT. Data collection took place during two 1-hour class periods separated by one week, with task order counterbalanced within each testing session. During the first data-collection session, participants completed paper-and-pencil versions of the VLT and G_Lex, before completing a background questionnaire at the end of the session. For the second session, participants responded to the PVLT and Lex30. Tests were divided between the two class sessions in such a way to allow for testing sessions of the same length.

3.3 *Data Coding and Analysis*

To allow for comparability across tests, we used one single frequency benchmark for scoring. For this purpose, we opted to use the online BNC-COCA corpora (Nation, 2017). On Lex30 and G_Lex, any response provided that was not a function word, proper noun, or number and did not belong to the first 1,000 most frequent words in the reference corpora received one point. Although the VLT and the PVLT are also intended to assess knowledge of words that lie outside of the first 1,000 frequency band, these tests target specific words which were chosen based on frequency information available when they were designed. This means that the frequency of certain target items may have changed and, importantly, some may now fall within the first 1,000 most frequent word families according to more recent frequency information. We thus verified the frequency of each target item in the BNC-COCA, observing that 11 of the original 150 items on the VLT and that 5 of the original 90 items on the PVLT now belong to the 1,000 most frequent word families. Given that these tests are intended to target less frequent words (i.e., words beyond the 1,000 frequency band), we excluded those items.

This dataset was analyzed with an exploratory factor analysis using R (RStudio Team, 2020). To begin, we verified that the four scores from the four vocabulary tests were normally

distributed. One set of scores – from the VLT – was found to be non-normally distributed, because of a significantly negative skew. For this reason, we subjected these scores to a reverse score transformation (see Field, Miles & Field, 2012, p. 192). They were then log-transformed and reversed back. The final set of scores showed no problems with normality. In running the factor analysis, we followed general recommendations from Field et al. (2012) and field-specific recommendations from Loewen and Gonulal (2015) and Plonsky and Gonulal (2015). The results section is organized following the step-by-step process described by Loewen and Gonulal. To begin, we assess the appropriateness of conducting a factor analysis on our dataset. We then move on to decisions regarding the factor extraction method, the factor retention criteria, and the factor rotation method. Once these steps have been presented, we present the results from the factor analysis.

4. Factor Analysis

4.1 *Factorability of this Dataset*

Descriptive statistics for this dataset are presented in Table 3, whereas the full correlation matrix is provided in Table 4. The appropriateness of conducting a factor analysis on this dataset was assessed in three ways, namely by running Bartlett's test, the Kaiser-Meyer-Olkin (KMO) measure, and by checking the determinant of the correlation matrix. Beginning with Bartlett's test of sphericity, the Chi-square value was 155.245 (df = 6), which was significant at the $p > .001$ level. This indicates that the correlations in this dataset are large enough for factor analysis. KMO is a measure of sampling adequacy. In the case of this dataset, the overall KMO value was .74, and ranged from .7 to .79 for individual variables. According to Field et al. (2012, p. 770), values between .7 and .8 are considered good, which further supports the conclusion that factor analysis is appropriate for this dataset. Finally, we explored whether there may be issues of multicollinearity in this dataset by calculating the

determinant of the correlation matrix. The determinant for this matrix was 0.2012485, which suggests that there are no problems with multicollinearity.

Table 3

Descriptive Statistics

Measure	Mean	Median	*SD*	range
Lex30	40.12	42	13.80	3-77
G_Lex	17.96	17.5	9.68	0-41
PVLT	47.01	49	11.39	18-71
VLT	104.31	112	23.87	28-135

Table 4

Correlation Matrix

	Lex30	G_Lex	PVLT	VLT
Lex30	1	0.569	.616	.482
G_Lex		1	.527	.356
PVLT			1	.689
VLT				1

4.2 *Factor Analysis Decisions*

The first decision to be made concerned the type of factor analysis to carry out. Because we were interested in uncovering potential latent constructs underlying the four variables measured (see Bandalos & Boehm-Kaufman, 2009, p. 63), exploratory factor analysis was considered more suitable than principal component analysis. To conduct the analysis in RStudio, we used the fa() command from the psych package with the factor

extraction method of minimum residual.[2] In order to determine the number of factors to extract from the dataset, we followed Loewen and Gonulal's (2015, pp. 196-197) recommendation and explored several factor retention criteria. We began by using parallel analysis (with the fa.parallel() command in the psych package). In this approach to factor retention, the eigenvalues of the actual variables are compared with randomly generated eigenvalues based on a data matrix of the same size as the original dataset. Only actual eigenvalues that are larger than the generated ones are recommended to be retained, regardless of their absolute value. In the case of this dataset, parallel analysis showed a two-factor solution, with the first factor having an eigenvalue of 2.30 and the second of .33. Thus, although the absolute eigenvalue associated with the second factor is below the recommended values of 1.0 (Kaiser's cut-off) and 0.7 (Joliffe's cut-off), parallel analysis suggests a two-factor solution. It should be noted that across-the-board cut-offs like Kaiser's and Joliffe's have come under criticism, with most empirical assessments of retention criteria showing parallel analysis to perform better than both (see Bandalos & Boehm-Kaufman, 2009; Field et al., 2012, p. 764). We next looked at cumulative variance. When using a two-factor solution, as suggested by parallel analysis, the first factor accounts for 57% and the second factor for 8% of variance, for a total cumulative variance of 65%. If we follow Loewen and Gonulal (2015, p. 194), who suggested that "it may be appropriate to continue factor extraction until at least 60% of the total variance is accounted for," a two-factor solution is justified. Finally, inspection of the scree plot revealed inflection points compatible with the retention of one or two factors. Taken together, we retained a two-factor solution for the final analysis. The final decision that we made at this stage concerned the method of factor rotation. As we expected that the two factors may correlate (see discussion in Plonsky & Gonulal, 2015, p. 22), we opted for oblique rotation (more specifically, oblimin).

[2] We also tested other extraction methods, such as maximum likelihood; these changes had no impact on the results.

4.3 *Factor Loadings*

The pattern matrix for the rotated factor loadings is provided in Table 5. Each of the four tests loaded strongly (> 0.3, see Loewen & Gonulal, 2015, p. 199) on only one factor. On the one hand, Lex30 and G_Lex loaded strongly on Factor 2, whereas the PVLT and VLT showed high loading values on Factor 1.

Table 5

Rotated Factor Loadings (Pattern Matrix)

	Factors	
Measure	1	2
Lex30		0.59
G_Lex		0.80
PVLT	0.69	
VLT	0.87	

5. Interpretation and Discussion

The interpretation of factor loadings requires that the analyst identify the core content for each factor (Loewen & Gonulal, 2015, p. 203). To aid in the identification of this core content for the two factors identified in our exploratory factor analysis, we returned to Table 2, in which we identified the different components of vocabulary knowledge called upon by each of the four tests under study (using Nation's, 2013, components approach to vocabulary knowledge). In Table 6, we have grouped together the tests as a function of their factor loadings (the PVLT and VLT on Factor 1, Lex30 and G_Lex on Factor 2). Under each test, we have reproduced the aspects of vocabulary knowledge potentially tapped by each (taken

from Table 2). In addition, we have bolded aspects of vocabulary knowledge that are shared by tests loading onto the same factor. In terms of core content for Factor 1, the PVLT and VLT overlap with regards to one aspect of vocabulary knowledge, namely receptive knowledge of the written form of a word. In addition, both of these tests are predicted to tap two other aspects of receptive vocabulary knowledge. For Factor 2, three cases of overlap are visible, all three of which involve productive vocabulary knowledge: productive knowledge of written form, productive knowledge of form and meaning, and productive knowledge of word associations. By considering the aspects of vocabulary knowledge solicited by each of the four tests, it appears that at the core of Factor 1 lies the assessment of receptive knowledge, whereas the thematic core of Factor 2 is the measurement of productive vocabulary knowledge. We thus suggest that Factor 1 corresponds to receptive vocabulary knowledge, whereas Factor 2 reflects productive vocabulary knowledge.

Table 6

Core Content for Factor Loadings

Factor 1 *Receptive knowledge*	
PVLT	VLT
Receptive written form Receptive grammatical functions Receptive collocations Productive written form Productive grammatical functions Productive constraints on use Productive form and meaning	**Receptive written form** Receptive form and meaning Receptive associations
Factor 2 *Productive knowledge*	
Lex30	G_Lex
Productive written form **Productive form and meaning** **Productive associations**	**Productive written form** **Productive form and meaning** **Productive associations** Receptive collocations

In what follows, we discuss in turn each of the two factors identified in this analysis before then reflecting more generally on what this study brings to the discussion of the

constructs underlying vocabulary knowledge in a L2. We conclude our discussion by identifying the limitations of the current project and offering suggestions for future research.

At first blush, the results from Factor 1 – what we have labeled "receptive vocabulary knowledge" – might appear surprising. This is because the findings for this factor reveal that performance on the PVLT patterned with performance on the VLT, even though the former is billed as a test of productive vocabulary knowledge, whereas the latter is intended to tap receptive knowledge. However, as already mentioned, despite its widespread use as a measure of productive vocabulary knowledge, certain researchers (i.e., Webb, 2005) have expressed reservations about the PVLT as a measure of productive knowledge, highlighting the fact that certain of its design features may also lead to the assessment of receptive knowledge. The results from the present study provide empirical evidence in support of these reservations. The fact that performance on the PVLT patterned significantly with performance on the VLT is particularly important, given that the PVLT has been used as a benchmark in past validation studies for new tests of productive vocabulary knowledge (like Lex30; see Fitzpatrick & Meara, 2004; Walters, 2012). More specifically, the PVLT has often been used as a point of comparison for assessments of concurrent validity in correlational studies. Indeed, Walters (2012) reported high correlations (p = .772) between results obtained on PVLT and Lex30, whereas Fitzpatrick and Meara (2004) reported a more moderate correlation (p = .504). In our own results, the PVLT and Lex30 results showed a correlation at p = .616 (see Table 4). Although these relatively high bivariate correlations are consistent with the conclusion that the two tests are measuring the same or an overlapping construct (and, thus, these results may be – and have been – taken as evidence of concurrent validity), the results from our exploratory factor analysis suggest a different story, namely one in which the PVLT patterns with the measure of receptive vocabulary knowledge (the VLT). With respect to the PVLT, these results thus call into question whether this test may be the best choice for concurrent

validity studies concerning the assessment of productive vocabulary knowledge. At a more general level, the divergence seen between validity arguments based on bivariate correlations and the factor analysis we have presented highlights the importance of varied and complementary approaches to assessing construct validity, as well as demonstrating the need for ongoing and sustained attention to questions of validity.

Turning now to Factor 2, Lex30 and G_Lex both loaded strongly onto this factor, suggesting that these two tests are tapping into the same construct. Fitzpatrick and Clenton (2017) created G_Lex to closely match Lex30 insofar as on both tests participants respond to several activation events (meaning elements that solicit vocabulary knowledge) and scores are based on the number of infrequent words provided. G_Lex crucially differs from Lex30 with respect to the presence of context (in the form of carrier sentences), a difference which Fitzpatrick and Clenton hypothesize will result in differences in the vocabulary knowledge reflected in performance (see their vocabulary test capture model). However, the fact that performance on these two tests patterns together in our analysis suggests that the hypothesized difference is more conceptual than empirical, at least for the current dataset. Considering the core content covered by these two tests, we named this factor "productive vocabulary knowledge." The fact that Lex30 and G_Lex load strongly onto this factor is consistent with the intended aim of these tests, which was to provide a general assessment of productive vocabulary knowledge. Moreover, whereas Lex30 has been the focus of numerous validation attempts, the more recently developed G_Lex has yet to receive such attention. Thus, one of the notable contributions of the present study is to provide initial evidence for the construct validity for G_Lex. The results with respect to this test are promising and will hopefully lead to additional research involving this measure.

The results of our exploratory factor analysis and our interpretation of the factor loadings thus suggest that the receptive versus productive knowledge distinction underlies the

primary constructs assessed by the four vocabulary tests under study. Our findings echo those of González-Fernández and Schmitt (2020), who reported that the most parsimonious SEM model of their data identified as crucial the distinction between recall (i.e., the ability to retrieve word knowledge) and recognition (i.e., the ability to recognize and select word knowledge, see p. 486). These results have potentially important implications for theory. In Nation's (2013) widely cited list of vocabulary knowledge components (see [1] and Table 2), vocabulary knowledge is first organized into nine components, with each component then covering receptive and productive knowledge. For example, the component "spoken form" encompasses the receptive knowledge of what the word sounds like and the productive knowledge of how to pronounce the word. As González-Fernández and Schmitt (2020, pp. 500-501) highlight, this presentation implies a hierarchical organization, with vocabulary knowledge components on the first tier and the receptive-productive knowledge distinction conceptualized within each component. However, both the results from the present study and those reported in González-Fernández and Schmitt (2020) offer evidence of the potentially greater importance of the receptive-productive distinction. Additional research is necessary, but these initial findings suggest that a distinction between the receptive and the productive knowledge constructs might prove to be the most fundamental division along which vocabulary knowledge in a L2 is organized.

To end our discussion, we identify two limitations of our study, which also serve to highlight potential directions for future research. The first limitation concerns the participant population who took part in this experiment. As English majors well on their way to a university degree in English studies, these participants were proficient in their L2. In the context of the present study, the relative homogeneity in L2 proficiency constitutes a potential limitation insofar as it has been suggested that certain assessments of vocabulary knowledge may tap different types of knowledge as a function of the learners' overall proficiency level.

For example, Walters (2012) administered Lex30 to Turkish learners of English at three proficiency levels. At a subsequent testing time, the participants also completed a sentence elicitation test 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 elicitation task 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 "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 light of these results, it is unclear whether our findings can be generalized to other proficiency levels, a possibility that should be explored in future research. A second limitation concerns the number of different tests used in this study. The main focus of our project was on the assessment of productive vocabulary knowledge. However, there currently exist few reliable tools to assess this construct (Miralpeix, 2020, p. 192). In addition to three published assessment tools of productive vocabulary knowledge (PVLT, Lex30, and G_Lex), we also decided to include one test of receptive vocabulary knowledge (the VLT), given doubts about the construct validity of the PVLT. This small number of measures meant that only two tests loaded on each factor, which is, according to Loewen and Gonulal (2015, p. 203), the minimum number of variables needed for a meaningful interpretation. Future research would thus do well to include a larger number of tests, which may allow to potentially confirm and strengthen the current results.

6. Conclusion

The impetus for this study was a desire to explore the construct validity of measures used to assess productive vocabulary knowledge in a L2. Vocabulary researchers have shown strong interest in questions of construct validity, which has been addressed using a variety of

analytic approaches. We contributed to this discussion with an exploratory factor analysis. Our results showed that for a group of 100 Francophone learners of L2 English, performance on the PVLT and the VLT patterned together (Factor 1), while performance on Lex30 and G_Lex loaded strongly onto the same factor (Factor 2). We interpreted Factor 1 as representing receptive vocabulary knowledge and Factor 2 as reflecting productive vocabulary knowledge. These results allowed us to offer insights into the nature of the four tests under study and additional evidence of the importance of the receptive-productive knowledge distinction in the conceptualization of L2 vocabulary.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford University Press.

Bandalos, D. L., & Boehm-Kaufman, M. R. (2009). Four common misconceptions in exploratory factor analysis. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 61-87). Routledge.

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

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

Clenton, J., de Jong, N. H., Clingwall, D., & Fraser, S. (2021). Investigating the extent to which vocabulary knowledge and skills can predict aspects of fluency or a small group of pre-intermediate Japanese L1 users of English (L2). In J. Clenton & P. Booth (Eds.), *Vocabulary and the four skills: Pedagogy, practice, and implications for teaching vocabulary* (pp. 126-145). Routledge.

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

Eguchi, M., Suzuki, S., & Suzuki, Y. (2021). Lexical competence underlying second language word association tasks: Examining the construct validity of response type and response time measures. *Studies in Second Language Acquisition*. Advance online publication. http://doi.org/10.1017/S0272263121000164.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*, 141-172. https://doi.org/10.1017/S0272263105050096.

Ellis, R., & Loewen, S. (2007). Confirming the operational definitions of explicit and implicit knowledge in Ellis (2005): Responding to Isemonger. *Studies in Second Language Acquisition*, *29*, 119-126. https://doi.org/10.1017/S0272263107070052.

Elmetaher, H. (2021). Investigating productive vocabulary knowledge development: A task-based approach. *Studies in European and American Cultures Journal, 28*, 1-25. http://doi.org/10.15027/52374.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. SAGE Publications.

Fitzpatrick, T. (2007). Productive vocabulary tests and the search for concurrent validity. In H. Daller, J. Milton & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary* (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*, 537-554. http://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 http://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.

Gilyuk, V., Edmonds, A., & Sneed German, E. (2021). Exploring the evolution in oral fluency and productive vocabulary knowledge during a stay abroad. *Journal of the European Second Language Association, 5*(1), 101-114. http://doi.org/10.22599/jesla.80.

González-Fernández, B., & Schmitt, N. (2020). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics, 41*(4), 481-505. https://doi.org/10.1093/applin/amy057.

Gutiérrez, X. (2013). The construct validity of grammaticality judgment tests as measures of implicit and explicit knowledge. *Studies in Second Language Acquisition, 35,* 423-449. https://doi.org/10.1017/S0272263113000041.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition, 21*(2), 303-317. https://doi.org/10.1017/s0272263199002089.

Langsford, S., Perfors, A., Hendrickson, A. T., Kennedy, L. A., & Navarro, D. J. (2018). Quantifying sentence acceptability measures: Reliability, bias, and variability. *Glossa: a Journal of General Linguistics, 3*(1), 1-34. https://doi.org/10.5334/gjgl.396.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics, 19*(2), 255-271. https://doi.org/10.1093/applin/19.2.255.

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

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

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

Loewen, S., & Gonulal, T. (2015). Exploratory factor analysis and principal components analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 182-212). Routledge.

Meara, P. (1997). Towards a new approach to modelling vocabulary acquisition. In N. Schmitt, & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 109-121). Cambridge University Press.

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., & Wolter, B. (2004). V_LINKS: Beyond vocabulary depth. In D. Albrechtsen, K. Haastrup, & B. Henriksen (Eds.), *Angles on the English speaking world 4* (pp. 85–96). Museum Tusculanum Press.

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

Milton, J. (2007). Lexical profiles, learning styles and the construct validity of lexical size tests. In J. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary* (pp. 47-58). Cambridge University Press.

Miralpeix, I. (2020). L1 and L2 vocabulary size and growth. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 189-206). Routledge.

Nation, P. (1983). Testing and teaching vocabulary. *Guidelines, 5,* 12–25.

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

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

Nemati, A. (2010). Active and passive vocabulary knowledge: The effect of years of instruction. *Asian EFL Journal*, *12*, 30-46.

Palmberg, R. (1987). Patterns of vocabulary cevelopment in foreign-language learners. *Studies in Second Language Acquisition*, *9*(2), 201-219. https://doi.org/10.1017/S0272263100000474

Plonsky, L., & Gonulal, T. (2015). Methodological synthesis in quantitative L2 research: A review of reviews and a case study of exploratory factor analysis. *Language Learning, 65*, Supp. 1, 9-35. https://doi.org/10.1111/lang.12111

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

Read, J. (2020). Key issues in measuring vocabulary knowledge. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 545-560). Routledge.

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC. http://www.rstudio.com/.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual.* Palgrave Press.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning, 64*(4), 913-951. https://doi.org/10.1111/lang.12077.

Schmitt, N. (2019). Understanding vocabulary acquisition, instruction, and assessment: A research agenda. *Language Teaching, 52*, 261-274. https://doi.org/10.1017/S0261444819000053.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55-88. https://doi.org/10.1177/026553220101800103.

Schoonen, R. (2015). Structural equation modeling in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 213-242). Routledge.

Tseng, W.-T., & Schmitt, N. (2008). Toward a model of motivated vocabulary learning: A structural equation modeling approach. *Language Learning*, *58*, 357-400. https://doi.org/10.1111/j.1467-9922.2008.00444.x.

Uchihara, T., Eguchi, M., Clenton, J., Kyle, K., & Saito, K. (2021). To what extent is collocation knowledge associated with oral proficiency? A corpus-based approach to word association. *Language and Speech*. Advance online publication. https://doi.org/10.1177/00238309211013865

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. (Ed.) (2020). *The Routledge handbook of vocabulary studies*. Routledge.

Webb, S. (2021). The lemma dilemma: How should words be operationalized in research and pedagogy? *Studies in Second Language Acquisition*, *43*(5), 941-949. https://doi.org/10.1017/S0272263121000784.

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL – International Journal of Applied Linguistics, 168*(1), 33-69. https://doi.org/10.1075/itl.168.1.02web.

Williams, J., Segalowitz, N., & Leclair, T. (2014). Estimating second language productive vocabulary size: A capture-recapture approach. *The Mental Lexicon, 9*(1), 23-47. https://doi.org/10.1075/ml.9.1.02wil.

Yanagisawa, A., & Webb, S. (2020). Measuring depth of vocabulary knowledge. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 371-386). Routledge.

Zhang, D. (2012). Vocabulary and grammatical knowledge in second language reading comprehension: A structural equation modeling study. *The Modern Language Journal*, *96*, 558-575. https://doi.org/10.1111/j.1540-4781.2012.01398.x.

Appendix

Figure 1A. Example of completed Lex30

Time: 15 minutes

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

1.	attack	heart	bandit	serious	sudden
2.	board	game	scholar	black	white
3.	close	down	door	far	friend
4.	cloth	beautiful	ugly	wet	small
5.	dig	tomb	sand	undertaker	death
6.	dirty	clean	mad	outside	rain
7.	disease	deadly	pandemy	clinic	doctor
8.	experience	amazing	unforgettable	together	improve
9.	fruit	good	green	vegetable	field
10.	furniture	school	holyday	sad	carpet
11.	habit	bad	common	hobby	good
12.	hold	on	light	let go	hand
13.	hope	dove	optimistic	pesimistic	olive
14.	kick	boxing	hurt	punch	bruise
15.	map	cart	country	road	orientation course race
16.	obey	army	dog	parents	law
17.	pot	a few	flower	dish	
18.	potato	Ireland	chips	fat	yellow
19.	real	fake	imagination	fantasy	boring
20.	rest	dead	food	trash	
21.	rice	Asia	food	white	rice field
22.	science	high-school	future	improvement	experience
23.	seat	amphitheatre	couch	lazy	obidience
24.	spell	magic	cast	wizard	Hogward
25.	substance	strange	mixture	cauldron	drug
26.	stupid	smart	other	person	
27.	television	news	tvshow	couch	parents
28.	tooth	brush	white	dentist	publicity
29.	trade	globalisation	market	store chain	
30.	window	sun	bird	view	cigarette

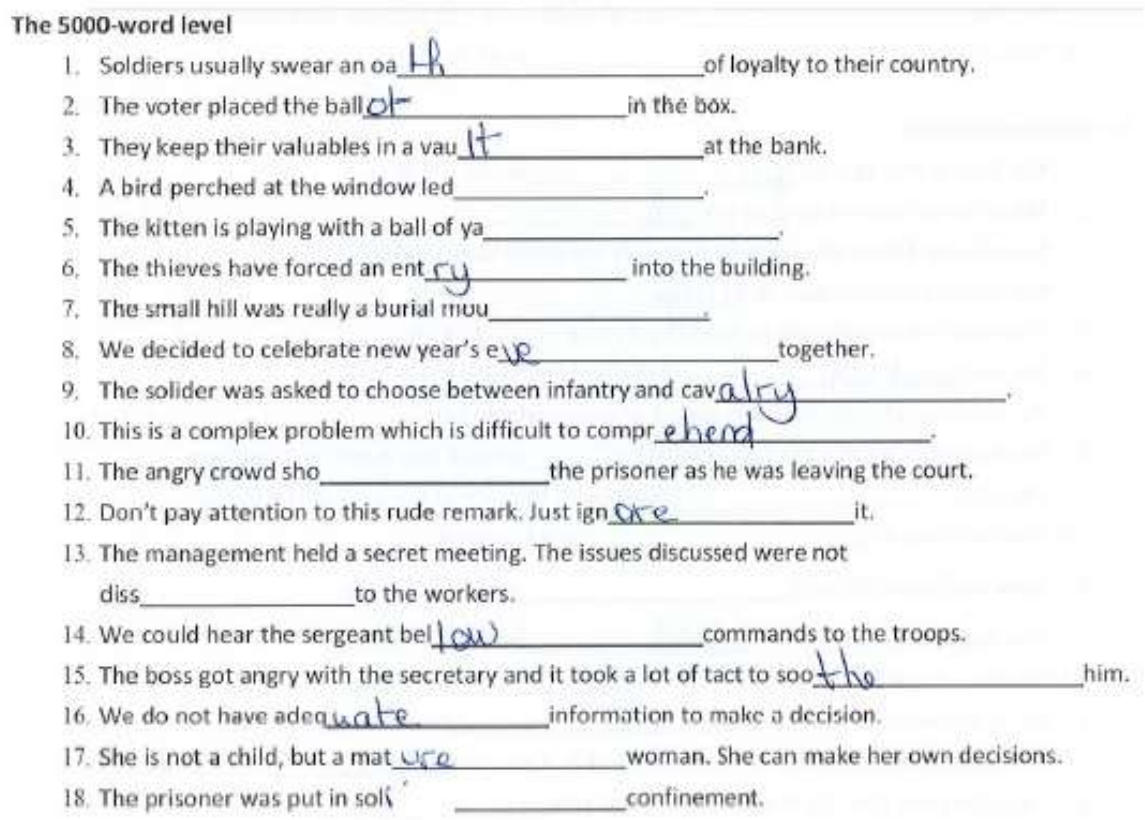
Figure 1B. Example of completed G_Lex

Time: 15 minutes

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

1. She loved to_____over the phone.	cry	laugh	yell		
2. When I feel sad I always go to the_____.	church	beach	bar	house of my parents	library
3. They think car-racing is_____.	funny	expensive	boring	dangerous	amazing
4. His colleague wanted to_____the report.	cancel	fill	save	burn	hide
5. My favourite_____is football.	game	sport			
6. She looked_____when she saw her friends.	happy	sad	suprised	worried	
7. He couldn't_____the car.	stop	open	leave	start	buy
8. With a fire in my house I would save my_____.	dog	computer	books	family	life
9. Many people feel_____about the environment.	were concerned	worried			
10. The parents_____the children.	of	with			
11. He was happy with his_____.	dog	friend	life	love	brother
12. He didn't think her teacher was_____at all.	busy	severe	cool		
13. She always wanted to_____after a busy day at work.	sleep	read	talk	walk	cry
14. She sent_____to her mother.	kisses	flowers	cards	mails	love
15. The weather looked_____before the game.	awful	good			
16. He wanted to_____the letter.	read	analyse	throw	buy	answer
17. She was excited about_____.	the trip	leaving	.		
18. The girls thought the rock concert was_____.	amazing	loud	annoying	an experience	a mess
19. He took the chance to_____the president.	call	see	touch	speak to	photographe
20. He gave his boss_____.	a reward	a letter			
21. At the funeral the family felt_____.	sorry	alone	relived	oppresed	
22. He always_____his breakfast.	swallow	eat	forget	skip	
23. She put the food in the_____.	trash	plate			
24. She was always_____to those who needed help.	nice	near	far	mean	

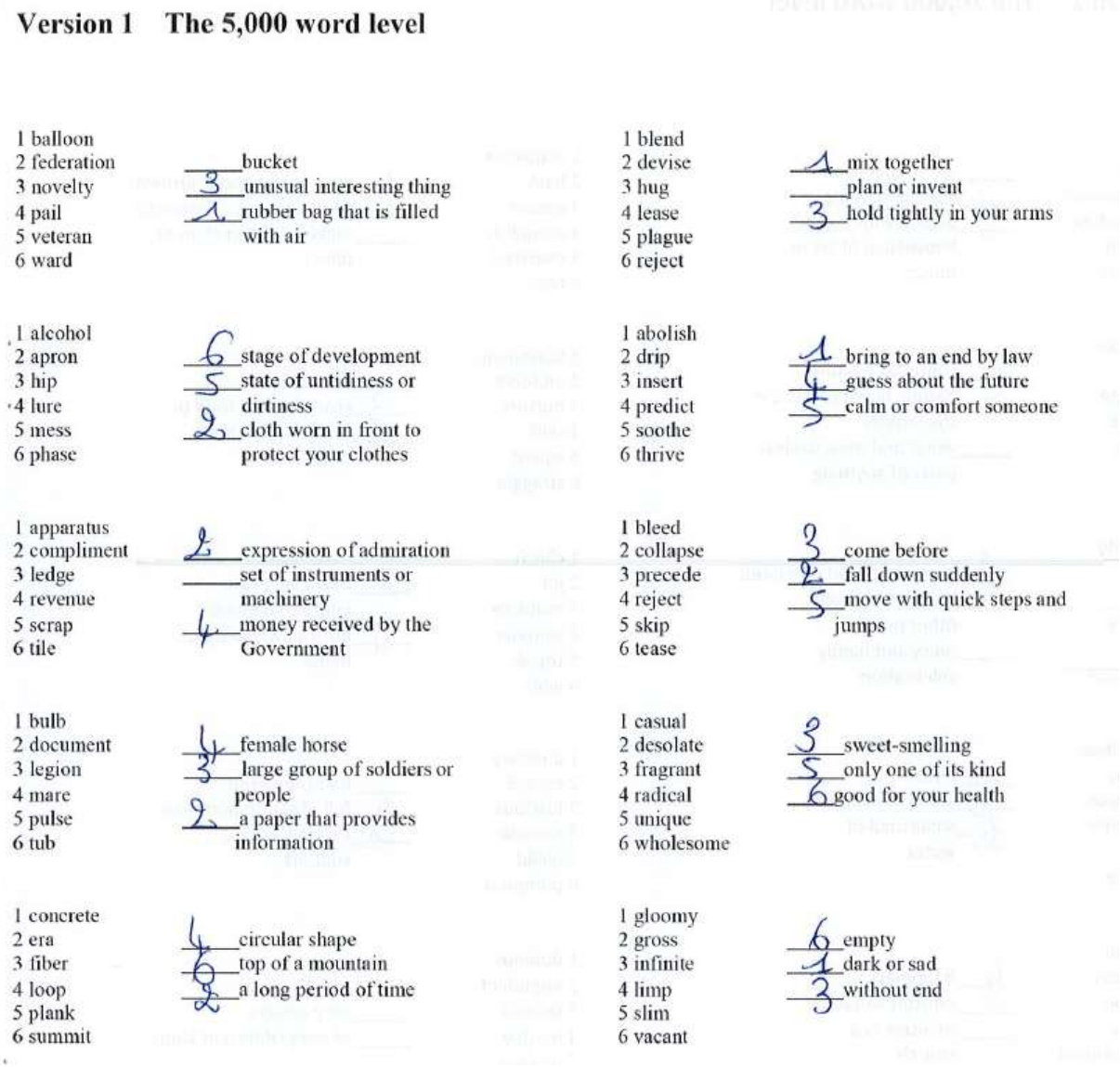
Figure 1C. Example of completed PVLT (5K band)



The 5000-word level

1. Soldiers usually swear an oa**th** of loyalty to their country.
2. The voter placed the ball**ot** in the box.
3. They keep their valuables in a vau**lt** at the bank.
4. A bird perched at the window led**ge**.
5. The kitten is playing with a ball of ya**rn**.
6. The thieves have forced an ent**ry** into the building.
7. The small hill was really a burial mou**nd**.
8. We decided to celebrate new year's e**ve** together.
9. The solider was asked to choose between infantry and cav**alry**.
10. This is a complex problem which is difficult to compr**ehend**.
11. The angry crowd sho**uted** 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**ed** to the workers.
14. We could hear the sergeant bel**low** commands to the troops.
15. The boss got angry with the secretary and it took a lot of tact to soo**the** him.
16. We do not have ade**quate** 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 sol**itary** confinement.

Figure 1D. Example of completed VLT (5K band only)



Version 1 The 5,000 word level

1 balloon		1 blend	
2 federation	___ bucket	2 devise	4 mix together
3 novelty	3 unusual interesting thing	3 hug	___ plan or invent
4 pail	1 rubber bag that is filled	4 lease	3 hold tightly in your arms
5 veteran	with air	5 plague	
6 ward		6 reject	
1 alcohol		1 abolish	
2 apron	6 stage of development	2 drip	1 bring to an end by law
3 hip	5 state of untidiness or	3 insert	4 guess about the future
4 lure	dirtiness	4 predict	5 calm or comfort someone
5 mess	2 cloth worn in front to	5 soothe	
6 phase	protect your clothes	6 thrive	
1 apparatus		1 bleed	
2 compliment	2 expression of admiration	2 collapse	3 come before
3 ledge	___ set of instruments or	3 precede	2 fall down suddenly
4 revenue	machinery	4 reject	5 move with quick steps and
5 scrap	4 money received by the	5 skip	jumps
6 tile	Government	6 tease	
1 bulb		1 casual	
2 document	4 female horse	2 desolate	3 sweet-smelling
3 legion	3 large group of soldiers or	3 fragrant	5 only one of its kind
4 mare	people	4 radical	6 good for your health
5 pulse	2 a paper that provides	5 unique	
6 tub	information	6 wholesome	
1 concrete		1 gloomy	
2 era	4 circular shape	2 gross	6 empty
3 fiber	6 top of a mountain	3 infinite	1 dark or sad
4 loop	2 a long period of time	4 limp	3 without end
5 plank		5 slim	
6 summit		6 vacant	