Investigating Productive Vocabulary Knowledge Development: A Task-Based Approach

Hosam ELMETAHER

Background

A number of studies have attempted to explore productive vocabulary development (e.g., Housen et al., 2008; Crossley et al., 2009, 2010; Berger et al., 2017). Housen et al. (2008) investigated the vocabulary development of a group of Dutch speaking learners of French (n = 19) in a three-year study. Their study explored L2 lexical proficiency with a variety of different measures to identify which measures of productive vocabulary capture lexical proficiency. They reported that their participants developed in terms of lexical diversity, lexical sophistication, and lexical productivity.

A further longitudinal study from Crossley et al. (2009) investigated vocabulary development in the spontaneous speaking of L2 learners. Crossley et al. 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 of six undergraduate learners at six different points over the course of one year. Crossley et al. 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. Interestingly, for most (five of their six participants), they reported a linear trend of lexical development (Crossley et al., 2009).

Crossley et al. (2010) conducted a year-long longitudinal study examining the vocabulary produced by six beginning undergraduate learners of English (L2), which was examined for polysemy, word senses, and word use frequency. Crossley et al. (2010) suggested that their study shows their participants' lexical development along with lexical network growth. They reported that participant use of polysemy increased with word use frequency from the 2nd to the 16th week of their observations. They also found that their participant group developed their word senses for a specific set of six words. Crossley et al. (2010) suggested that their findings, when taken together, indicate that lexical development relates to learner use of polysemy and a related extension of core meanings of polysemous

words, and is, therefore, evidence of the growth of lexical networks.

Berger et al. (2017) reported on a longitudinal study undertaken in order to better understand how learner lexicons develop. Their paper comprises two studies: a cross-sectional study and a longitudinal study that I report here. Their longitudinal study analyzed spoken data from L2 adult learners over a year-long "study abroad" period. Berger et al. (2017) reported that their participants developed lexical proficiency over time, with longer time spent in English-speaking environments, leading to a greater production of more frequent words.

While the above studies describe productive lexical development, their findings are based exclusively on a performance-based approach and do not appear to capture how productive vocabulary knowledge develops in its various forms because of operationalization issues. For instance, Housen et al. (2008) reported gains in lexical diversity, sophistication, and productivity from a variety of different measures but contend that "each type of measure and each operationalisation has its own inherent strengths and problems" (p. 281). Similarly, Crossley et al. (2009) did not show a strong relationship between hypernymic development and lexical proficiency because their study depends on measuring lexical diversity, which is "not completely representative of lexical proficiency" (p. 329). Similarly, Crossley et al. (2010) reported that their mixed methods approach, which adopted a quantitative then a qualitative approach, might produce different results with infrequent items such as those used by Schmitt (1998). Moreover, despite Berger et al. (2017) demonstrating the development of L2 spoken proficiency over time, they suggested their findings are limited because their approach considered the lexicon as consisting of single words and that future development studies should include analysis of a potential increased use of multi-word units in relation to increases in lexical proficiency. Such operationalization concerns issue a call for a productive vocabulary knowledge development study that concurrently uses measures without inherent problems, is based on research representative of the measure, and employs an analysis of infrequent items. The design of the current paper is, therefore, intended to respond to this call. In the next section, I attempt to define the construct of productive vocabulary knowledge and some of its measures.

Productive Vocabulary Knowledge Construct

Emerging consensus suggests that eliciting productive vocabulary knowledge is far from a straightforward endeavor (e.g., Clenton, 2010; Fitzpatrick & Clenton, 2010; Fitzpatrick, 2007; Fitzpatrick & Clenton, 2017; Milton, 2009; Walters, 2012). Methods of assessment vary and tend to be based on items carefully selected according to frequency (Ellis, 2002), specific contextual concerns, or specific task demands. Tests based on frequency counts (i.e., by comparing L2 participant data with online corpora organized according to relative L1 frequency usage) are problematic for several reasons, not least because frequency lists may not comfortably relate to items produced in writing or speaking (Milton, 2009). Based on Zipf's law (1936), the relative frequency of a word relates to its rank in a frequency table, such that words ranked first in the table occur twice as frequently as those ranked second, and those ranked second in the table occur twice as frequently as those ranked third, etc. That this relationship is "not perfect" (Milton, 2009, p. 45) is a concern, but does not deter researchers (e.g., Nation, 2001) from suggesting that learners should do everything possible to learn the first 2,000 most frequent words. Others (e.g., Kremmel, 2016) suggest that dividing words into 1,000-item frequency bands is far from ideal, likely "arbitrary," and potentially "variable" in terms of the degree of "clustering power of frequency" (p. 980). A further confounding factor in eliciting productive vocabulary knowledge relates to contextual influences because context influences what learners can produce in response to different task formats such as in the form of single words, sentences, or compositions- thus eliciting the need to adhere to composition genre (e.g., demanding knowledge of a variety of composition genres such as opinion, description, narration, etc.). That tests might vary also relates to other factors such as test type (e.g., entrance test, diagnostic test). Such different task demands are the concern of the current paper. From this point, I turn to presenting a review of vocabulary tasks (e.g., LFP: Laufer & Nation, 1995; PVLT: Laufer & Nation, 1999; Lex30: Meara & Fitzpatrick, 2000; and G_Lex: Fitzpatrick & Clenton, 2017) in an attempt to highlight the different task demands and the potential for influences on performance outcomes.

The Lexical Frequency Profile (LFP)

The Lexical Frequency Profile (LFP) (Laufer & Nation, 1995) is a composition task that also examines learner output according to frequency. The LFP requires participants to write two paragraphs of about 300–350 words each over two different class periods (e.g., "*Should a government be allowed to limit the number of children a family can have? Discuss*

this idea considering basic human rights and the danger of population explosion"). Participant compositions are then processed according to four criteria: (i) incorrectly used words are excluded; (ii) misspellings are corrected; (iii) approximately written words from a word family are tolerated; and, (iv) all proper nouns are excluded from the analysis. For the standard LFP, data are ordinarily categorized as belonging to one of four frequency bands (1k, 2k, the University Word List [UWL], and those words not in the list [nil]).

The Productive Vocabulary Levels Test (PVLT)

The Productive Vocabulary Levels Test (PVLT) (Laufer & Nation, 1999; Figure 1) is a "controlled productive" word completion task in which incomplete words are presented in sample sentences. The task has been widely used and is the focus of several papers (e.g., Fitzpatrick, 2007; Fitzpatrick & Clenton, 2010; Henriksen & Danelund, 2015; Laufer, 1998; Laufer & Nation, 1995, 1999; Laufer & Paribakht, 1998; Meara & Alcoy, 2010; Read, 2012; Stæhr, 2009; Walters, 2012; Webb, 2009). For the PVLT, the first few letters of each word 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, 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), "mastery" of one level is likely a "matter of judgement [..and] is probably around 15 or 16 out of 18 (85% or 90%)" (p. 41). Laufer and Nation also suggested that use of the PVLT enables researchers to investigate the "developments [that] occur in the different types of vocabulary over a period of time" (p. 45) and defined PVLT scores as "the total score of correctly retrieved items" (p. 39).

Figure 1 ***PVLT (Laufer & Nation, 1999, p. 46)***

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_____.

Lex30

Lex30 (Meara & Fitzpatrick, 2000; Figure 2) is "basically a word association task [reportedly] less constrained [than] context-limited productive tasks" (2000, p. 22). Lex30 has been widely used and appears in several recent research papers (e.g., 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). Carefully selected based on three criteria, each stimulus word (i) is highly frequent and from Nation's (1983) 1,000 frequency-level words; (ii) does not elicit a primary response in comparison with first language (English) speaker data (Edinburgh Associative Thesaurus; Kiss et al., 1973); and (iii) does not elicit common responses in comparison with first language (English) speaker responses (Edinburgh Associative Thesaurus; Kiss et al., 1973). Completed Lex30 task papers are typed and lemmatized. Individual corpora are then compared with online frequency counts 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 infrequent items.

Figure 2 *Lex30 (Meara & Fitzpatrick, 2000, pp. 28–29)*

1	attack				
2	board				
3	close				

G_Lex

G_Lex (Fitzpatrick & Clenton, 2017; Figure 3) is a sentence completion task in which up to 5 words are required to complete each of 24 sentence gaps (totaling a maximum of 120, the same as for the Lex30 task). Any words are accepted as long as the items fit the gap provided, with no hints or prompts designed to elicit knowledge of specific target words (and therefore different from the PVLT in this regard). The 24 sentences were designed to elicit an equal number (8) of nouns, adjectives, and verbs, to distinguish between different test takers in terms of their "(productive) lexical resource" (p. 855). Designed to meet five criteria, sentences: (i) are syntactically simple; (ii) contain only high frequency words; (iii) readily elicit five responses when trialed with first language (English) speakers; (iv) do not elicit lexical sets (e.g., *brown, blue, red*); and (v) do not elicit similar words in different sentences. G_Lex scoring is conducted in the same way as Lex30.

Figure 3 *G_Lex (Fitzpatrick & Clenton, 2017, p. 856)*

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

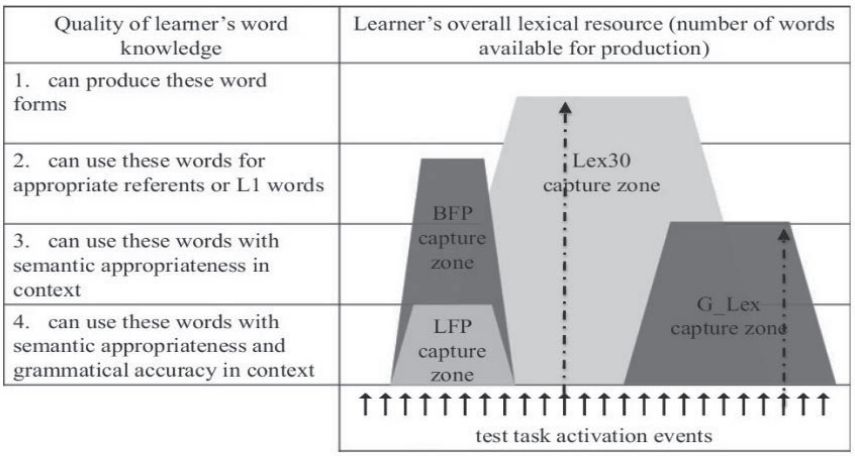
Comparisons between Productive Vocabulary Knowledge Tasks

In a comparison between different productive vocabulary tasks, Fitzpatrick and Clenton (2017) attempted to investigate whether productive vocabulary knowledge tasks are assessing similar knowledge. They found a significant correlation (r = .504, p < .01) between the PVLT and Lex30 task scores from which they suggest that the two "are assessing broadly similar constructs" (p. 545). In their attempts to interpret learner performances on tests of productive vocabulary, Fitzpatrick and Clenton (2017) presented three studies, beginning with a comparison of Lex30 (designed to elicit productive vocabulary knowledge with a word association task) with the LFP (a composition task). They reported that the correlation between the two task scores is not significant (r = .186, p = .098) and explored potential reasons, one relating to the two tasks' definition of "infrequent." They adjusted LFP scores to reflect "infrequent" in the same way as Lex30, but observed that correlations remain insignificant (r = .108, p = .339). Fitzpatrick and Clenton (2017) then highlighted the different task demands between the two (Lex30 and the LFP) tasks and devised a new (Brainstorm Frequency Profile (BFP)) task designed to explore the extent to which composition writing demands might impact task performance. Their BFP retains the LFP question task, but elicits responses in the form of single words as does Lex30. The authors found nonsignificant correlations between the Lex30 and BFP task scores (r = .153, p = .175). For their third study, Fitzpatrick and Clenton (2017) introduced the 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 in their BFP task. Their G_Lex task retains similarity to the LFP in the sense that the task requires test takers to respond to context. A comparison of G_Lex and Lex30 tasks is significant (r = .645, p < .01) and suggests that performance on one task is broadly predictive of the other.

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 (2017) devised a "Vocabulary Test Capture Model" (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 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 while the horizontal axis maps the quantity or breadth (see Figure 4). To demonstrate how users should interpret their model, Fitzpatrick and Clenton 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 Vocabulary 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 4 *Vocabulary Test Capture Model: Lex30, LFP, BFP, and G_Lex (Fitzpatrick & Clenton, 2017, p. 862)*



Experiment 1: Adding the Productive Vocabulary Levels Test to Fitzpatrick and Clenton's (2017) "Vocabulary Test Capture Model"

Fitzpatrick and Clenton (2017) suggest that the different axes of their model need investigating further regarding conceptualization of productive vocabulary knowledge. I therefore introduce an experiment in which I compare a widely cited vocabulary measure (the Productive Vocabulary Levels Test (PVLT); Laufer & Nation, 1995; 1999) with two tasks from Fitzpatrick and Clenton (2017): Lex30, and G_Lex.

Participants

I assessed 162 (86 male, 76 female) L1 Japanese second language (L2) undergraduate learners (of English) from two universities in western Japan. All participants were in their first year at university, aged 18, and had received regular English tuition from the age of 13 for approximately three hours per week. The participants were reported to be from a pre-intermediate to upper intermediate level of proficiency by their classroom teachers.

Procedure

Testing was conducted over a three-week period, with a week-long interval between each test. Classes comprised students from different faculties 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) were presented in different orders for each classroom group. The individual tasks were scored according to their original scoring procedures. Responses to both G_Lex and Lex30 were processed with JACET8000 word lists (Ishikawa et al., 2003) in order to better reflect the learner experience of the participants. Lex30 and G_Lex task scores consist of a count of all but highly frequent (i.e., 1k) responses, whereas the PVLT awards one point for each vocabulary item (out of 18 total) correctly provided at five frequency levels (2k, 3k, 5k, UWL, and 10k).

Results and Discussion

Table 1 shows the descriptive statistics for the three different productive vocabulary task scores. Scores appear to vary according to the task: Lex30 score is > G_Lex > PVLT. Table 2 shows a significant correlation between the three tasks ($p < .001$). The correlation between Lex30 and PVLT scores ($r = .599$, $p < .001$) is similar to that reported in the earlier comparison by Fitzpatrick and Clenton (2010) ($r = .504$, $p < .001$) and suggestive that the two tasks are "assessing broadly similar constructs" (p. 545). The correlation between Lex30 and G_Lex scores is also significant ($r = .503$, $p < .001$) but its effect size was not as strong as that reported in Fitzpatrick and Clenton (2010) ($r = .645$, $p < .001$), and might relate to differences in language proficiency. The correlation between G_Lex and PVLT scores was significant ($r = .400$, $p < .001$), suggesting that performance on one task is broadly predictive of the other.

Table 1 *Mean Scores, Standard Deviations (SD), and Range of Task Scores*

Score (N =162)	Mean Score	SD	Range
Lex30	24.85	11.74	1–52
G_Lex	17.15	6.69	1–39
PVLT	15.58	6.30	3–35

Table 2 *Correlations between the Three Productive Tasks (Lex30, G_Lex, and PVLT)*

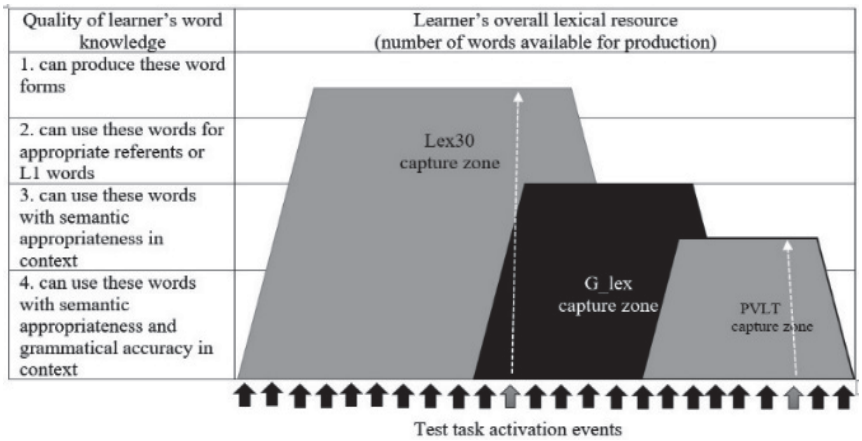
	Lex30	*G_Lex*
Lex30		
G_Lex	0.503***	
PVLT	0.599***	0.400***

*** $p < .001$

Differences in mean scores (Table 1), on the other hand, can perhaps more readily be explained with a revised Vocabulary Test Capture Model (Figure 5). For instance, only minimal quality of 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. Participant responses to the PVLT might, as Fitzpatrick and Clenton (2017) intimate, demonstrate a quality of word knowledge representative of a developed or developing lexicon.

I suggest that beginner level learners, with knowledge of predominantly high-frequency items, respond to vocabulary tasks with a similar proportion of infrequent items. As Daller et al. (2007) have suggested, learners at lower levels might share similar developmental trajectories to some extent and might, therefore, be predictable; however, as individual learners develop their vocabularies, an inconsistent pattern emerges of development in advanced learner lexicons.

Figure 5 *Revised Fitzpatrick and Clenton's "Vocabulary Test Capture Model" to Include Laufer and Nation's (1999) "Productive Vocabulary Levels Test"*



Experiment Two: Productive Vocabulary Knowledge Development: A Task-based Approach

With reference to my review of productive vocabulary development studies, this study intends to (i) employ measures without inherent problems, (ii) base studies on research that is representative of the measure, and (iii) employ an analysis of infrequent items. Accordingly, the design of my second experiment is intended to respond to these three concerns.

I also explore six other themes from the broader vocabulary literature that relate to: (i) the need to employ a multifaceted approach to my vocabulary knowledge development study on the basis that research (e.g., Schmitt, 1998) has shown the importance of a multifaceted view of vocabulary development; (ii) investigating the extent to which vocabulary knowledge develops differently according to different frequency bands. For example, Zhang and Lu (2013) indicated that different vocabulary constructs (in their study, breadth, and fluency) develop differently according to different frequency bands); (iii) the extent to which productive vocabulary knowledge follows a predictable developmental pattern (Daller et al., 2007); (iv) whether responses to Productive Vocabulary Levels Test enable researchers to investigate "developments [that] occur in the different types of vocabulary over a period of time" (Laufer & Nation, 1999, p. 45); (v) the extent to which vocabulary task influences performance (Fitzpatrick & Clenton, 2017); and (vi) the dynamics of the relationship between receptive vocabulary knowledge (i.e., needed for listening or

reading) and productive vocabulary knowledge (i.e., needed for writing and speaking) over time, which I believe has not yet been investigated (see Webb (2008) for a cross-sectional study examining the receptive–productive relationship).

I therefore include a receptive vocabulary knowledge task (XY_Lex; Meara & Miralpeix, 2016). X_Lex and Y_Lex have been widely used, and appear in several research papers (e.g., De Jong & Mora, 2017; Gilabert et al., 2009; Gyllstad & Wolter, 2016; Vasylets, Gilabert, & Manchón, 2017). X_Lex and Y_Lex require participants to respond to a yes/no task in which 120 words are presented and the answers to which indicate knowledge, respectively, of the 1,000–5,000 and 6,000–10,000 frequency bands. The tasks include several pseudowords and XY_Lex adjusts scores when such words are identified as genuine words. Accordingly, my research questions are:

1. Does performance on productive vocabulary tasks vary according to task and development?
2. Does performance on productive vocabulary tasks vary according to individual frequency bands?
3. Does performance on productive vocabulary tasks vary according to L2 proficiency?

Participants

I assessed 100 (60 male, 40 female) Japanese learners of (L2) English. All participants were in their first year of university, aged 18, and were from a wide range of university faculties. All had received regular English tuition from age 13, averaging three hours per week. The participants were reported to be from a pre-intermediate level of English language (L2) proficiency according to their classroom teachers.

Procedure

The second study employs the same three productive vocabulary tasks from the first study (Lex30, G_Lex, and the PVLT). Participants also completed a receptive vocabulary knowledge (yes/no) task (XY_Lex; Meara & Miralpeix, 2017). I conducted testing at three test points in one academic year (at 0, 6, and 9 months). I base the number of testing times on comments from Pellicer-Sánchez (2018) and Schmitt (2010) who both suggest that longitudinal testing more than two times might ease potential "typical pre-post-test design" result limitations. I also adopt a single academic year testing period based on Dóczi and Kormos' (2016) suggestion that "(t)he longer and greater the engagement, the more sizable the growth in word knowledge might be" (pp. 178–179).

Testing was conducted with four different class groups at each test time. I presented the tasks to the participants in different orders at each test time and compared results with the other test groups to negate test effects. As in the first experiment, the individual tasks were scored in accordance with their original scoring procedures. Responses to both G_Lex and Lex30 were processed with JACET8000 word lists (Ishikawa et al., 2003). Lex30 and G_Lex task scores consist of a count of all but highly frequent (i.e., 1k) responses, whereas the PVLT awards one point for each vocabulary item (out of 18 total) correctly provided at five frequency levels (2k, 3k, 5k, UWL, and 10k).

Because the second experiment is based on three test times, three different equivalent versions of each task were needed. 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 1) 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 new versions of G_Lex (B and C, Appendix 2) 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.

Results and Discussion

RQ1: Does performance on productive vocabulary tasks vary according to task and development?

Table 3 shows the different productive vocabulary task scores at the three different time points (Time 1, Time 2, and Time 3).

Table 3 *Productive Vocabulary Task Scores at the Three Different Times Points*

Time	Mean score		
	Lex30	G_Lex	PVLT
Mean scores at each time point			
T1	20.64 (8.93)	15.66 (6.08)	11.54 (5.59)
T2	21.20 (11.55)	20.49 (9.12)	11.61 (5.94)
T3	28.75 (12.21)	21.75 (8.74)	13.79 (6.79)
Changes between time points			
T2-T1	0.56	4.83***	0.07
T3-T2	7.55**	1.26	2.18***
T3-T1	8.11***	6.09***	2.25***

Note: ** $p < .01$, *** $p < .001$. Numbers in parentheses are standard deviations.

Each of the task scores appears to indicate productive vocabulary development although performance varies according to task as well as time. Lex30 data show a significant increase in scores from Time 2 to Time 3 and from Time 1 to Time 3. G_Lex data show a significant increase in scores from Time 1 to Time 2 and from Time 1 to Time 3. PVLT data show a significant increase in scores from Time 2 to Time 3 and from Time 1 to Time 3. Such findings highlight the dynamic nature of productive vocabulary growth.

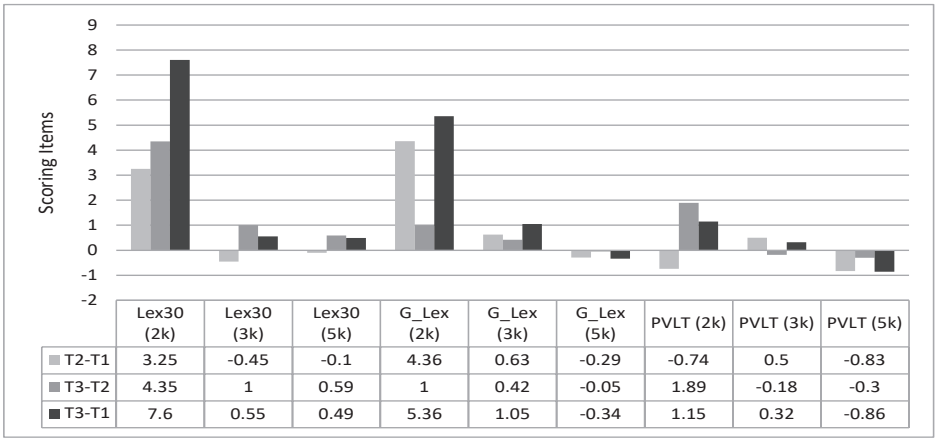
RQ2: Does performance on productive vocabulary tasks vary according to individual frequency bands?

To examine potential development, I report on words produced in the 2,000–, 3,000–, and 5,000-level frequency bands for all three tasks at the three test times. All participants produced words within these three frequency bands for all three productive vocabulary tasks (Lex30 and G_Lex tasks collects data from the 1k, 2k, 3k, 4k, 5k, 6k, 7k, and 8k frequency bands, and the PVLT collects participant knowledge from the 2k, 3k, 5k, 10k, and UWL frequency bands). I included all levels in my analysis, but do not report all of the results because the tasks did not elicit sufficient words at specific frequency levels (e.g., 8k, 10k, and UWL). The tasks share production of items within the 2k, 3k, and 5k frequency bands, and I report on task comparisons according to words produced below. Table 4 and Figure 6 show the different productive vocabulary task scores at the three different time points (Time 1, Time 2, and Time 3) according to individual frequency bands (2k, 3k, and 5k).

Table 4 *Productive Vocabulary Task Scores for 2,000, 3,000, and 5,000 Frequency Levels*

Time	Mean score								
	Lex30			G_Lex			PVLT		
	2,000	3,000	5,000	2,000	3,000	5,000	2,000	3,000	5,000
Mean scores at each time point									
T1	8.21 (4.54)	4.56 (2.53)	0.97 (0.89)	7.07 (3.39)	2.91 (1.83)	1.56 (0.92)	6.55 (2.75)	2.30 (1.67)	1.55 (1.49)
T2	11.46 (6.11)	4.11 (2.87)	0.87 (1.01)	11.43 (5.39)	3.54 (2.64)	1.27 (1.02)	5.81 (2.66)	2.80 (1.46)	0.70 (1.13)
T3	15.81 (6.73)	5.11 (3.19)	1.46 (1.25)	12.43 (4.40)	3.96 (2.74)	1.22 (0.90)	7.70 (3.17)	2.62 (1.66)	0.67 (0.92)
Changes between time points									
T2-T1	3.25**	−0.45	−0.10	4.36***	0.63	−0.29	−0.74**	0.50**	−0.85***
T3-T2	4.35***	1.00**	0.59***	1.00	0.42	−0.05	1.89***	−0.18	−0.03
T3-T1	7.60***	0.55	0.49***	5.36***	1.05***	−0.34**	1.15***	0.32	−0.88***

Note: ** $p < .01$, *** $p < .001$. Numbers in parentheses are standard deviations.



	Lex30 (2k)	Lex30 (3k)	Lex30 (5k)	G_Lex (2k)	G_Lex (3k)	G_Lex (5k)	PVLT (2k)	PVLT (3k)	PVLT (5k)
T2-T1	3.25	-0.45	-0.1	4.36	0.63	-0.29	-0.74	0.5	-0.83
T3-T2	4.35	1	0.59	1	0.42	-0.05	1.89	-0.18	-0.3
T3-T1	7.6	0.55	0.49	5.36	1.05	-0.34	1.15	0.32	-0.86

Figure 6 *Three Productive Vocabulary Task Scores for 2,000, 3,000, and 5,000 Frequency Levels Changes over 3 Test Times (Time 1, 2, and 3)*

For Lex30, items produced in the 2k band significantly increased over Times 1 to Time 2, Time 2 to Time 3, and Time 1 to Time 3; items produced in the 3k band significantly increased from Time 2 to Time 3; and items produced in the 5k band significantly increased from Time 2 to 3 and Time 1 to Time 3. For G_Lex, items produced in the 2k band significantly increased from Time 1 to Time 2 and from Time 1 to Time 3; items produced in the 3k band significantly increased from Time 1 to Time 3; and items produced in the 5k band significantly increased from Time 1 to Time 3. For the PVLT,

items produced in the 2k band significantly increased from Time 2 to Time 3 and from Time 1 to Time 3, but significantly decreased from Time 1 to Time 2; items produced in the 3k band significantly increased from Time 1 to Time 2; and items produced in the 5k band significantly decreased from Time 1 to Time 2 and from Time 1 to Time 3.

Taken together, these results might indicate a general improvement in participants' productive vocabulary knowledge. However, when I examine specific frequency levels for the three tasks, I observe decreases (i.e., Lex30 3k, 5k; G_Lex 5k; and PVLT 2k, 3k, and 5k). Such findings highlight the inconsistent developmental trajectory across frequency bands.

With even closer examination of the results, I observe significant positive change in Lex30 2k scores over the three testing times. Such significant positive change is not observed in the other two tasks (G_Lex and PVLT) or the other frequency bands (3k and 5k). I wonder whether such inconsistent changes in the 2k among the three tasks might relate to task characteristics as indicated in the revised Vocabulary Capture Model (Figure 5). To restate, Lex30 has the potential to elicit all levels of partial word knowledge, namely, knowledge restricted to form or to meaning. G_Lex has the potential to elicit an understanding of semantic appropriateness. The PVLT has the potential to elicit syntactic as well as semantic aspects of knowledge.

RQ3: Does performance on productive vocabulary tasks vary according to proficiency?

For my third research question, I first ran Pearson correlation analyses, comparing each of productive vocabulary tasks at the different test times with the receptive vocabulary task (XY_Lex). Table 5 shows that the relationships between receptive vocabulary size (XY_Lex) and the three productive vocabulary measures remained medium to large over time but inconsistently, suggesting that receptive and productive knowledge might tap different constructs.

Table 5 ***Pearson Correlations between the Three Productive Vocabulary Task Scores and XY_Lex Task Scores at the Three Different Test Times***

	Lex30	G_Lex	PVLT
Time 1	0.505***	0.356***	0.608***
Time 2	0.460***	0.496***	0.610***
Time 3	0.507***	0.448***	0.610***

Note: ***$p < .001$.

Both the PVLT and XY_Lex scores show consistent strong relationship (Time 1: r = 0.608, p < .001; Time 2: r = 0.610, p < .001; Time 3: r = 0.610, p < .001). At an early stage of development, what participants can recognize might also relate more closely to items that 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 the word). However, such a consistent strong relationship is not the same for the Lex30 and G_Lex task score comparisons. The strength of the relationship between receptive vocabulary and Lex30/G_Lex is moderate: Lex30 (Time 1: r = 0.505, p < .001; Time 2: r = 0.460, p < .001; Time 3: r = 0.507, p < .001); and, G_Lex (Time 1: r = 0.356, p < .001; Time 2: r = 0.496, p < .001; Time 3: r = 0.448, p < .001). Such differences might relate to the different task formats between Lex30/G_Lex and the PVLT with the former being more like free production and the latter requiring more control (Read, 2000). This result might indicate that the productive vocabulary measured by Lex30 and G_Lex, (i.e., the construct more aligned to the ability to "use" words compared to PVLT) might, to some extent, be independent of receptive vocabulary growth. This finding is potentially important because of the widely held assumption that there is always a strong relationship between receptive and productive vocabulary (e.g., Meara & Fitzpatrick, 2000; Webb, 2008); however, the evidence presented here suggests that the relationship might depend on productive (and receptive) vocabulary measures.

To examine whether productive vocabulary knowledge differed among participants with different L2 proficiency (receptive vocabulary sizes), I used a two-step cluster analysis with Schwarz's Bayesian Criterion to classify participants into groups with significantly different overall combined XY_Lex scores. The cluster analysis method is regarded as more objective and reliable than the use of judgements in grouping participants (e.g., median split). The cluster analysis produced three groups of participants with significantly different overall scores for each of the three time points (Table 6).

Table 6 *Three Productive Vocabulary Task Scores at the Three Different Time Points According to Different Proficiency Levels*

Group	*N*	Lex30	G_Lex	PVLT
Time 1				
Beginner	23	14.43** (8.53)	11.65** (6.04)	6.78** (4.13)
Intermediate	50	20.72** (7.98)	16.42 (5.55)	11.48** (4.21)
Advanced	27	25.77** (7.74)	17.66 (5.65)	15.70** (5.76)
Time 2				
Beginner	29	16.13 (7.55)	15.27 (6.53)	7.17** (3.69)
Intermediate	44	19.04 (9.24)	19.68 (8.39)	11.59** (4.57)
Advanced	27	30.14** (13.60)	27.40** (8.56)	16.40** (6.24)
Time 3				
Beginner	32	21.81** (10.81)	17.00** (6.00)	8.78** (4.81)
Intermediate	48	29.81** (10.99)	23.22 (8.69)	15.08 (5.31)
Advanced	20	37.30** (11.23)	25.80 (9.56)	18.70 (7.76)

Note: ** $p < .01$, *** $p < .001$. Numbers in parentheses are standard deviations.

At Time 1, aside from the difference between intermediate and advanced G_Lex task scores, all tasks showed significant differences in productive vocabulary knowledge (G_Lex, Lex30, PVLT) across the three proficiency groups: G_Lex [$F(2, 97) = 7.81$, $p < .01$, Beginner < Intermediate < Advanced], Lex30 [$F(2, 97) = 12.33$, $p < .01$, Beginner < Intermediate < Advanced], and PVLT [$F(2, 97) = 22.72$, $p < .01$, Beginner < Intermediate < Advanced]. At Time 2, aside from the differences between beginner and intermediate on Lex30 and G_Lex task scores, all tasks showed significant differences in productive vocabulary knowledge across the three proficiency groups: G_Lex [$F(2, 97) = 16.68$, $p < .01$, Beginner < Intermediate < Advanced], Lex30 [$F(2, 97) = 14.94$, $p < .01$, Beginner < Intermediate < Advanced], and PVLT [$F(2, 97) = 25.19$, $p < .01$, Beginner < Intermediate < Advanced]. At Time 3, aside from the differences between intermediate and advanced on G_Lex and PVLT, all tasks showed significant differences in productive vocabulary knowledge across the three proficiency groups: G_Lex [$F(2, 97) = 8.73$, $p < .01$, Beginner < Intermediate < Advanced], Lex30 [$F(2, 97) = 12.66$, $p < .01$, Beginner < Intermediate < Advanced], and PVLT [$F(2, 97) = 20.73$, $p < .01$, Beginner < Intermediate < Advanced].

Summary and Concluding Discussion

This current paper was designed to respond to an apparent gap in productive vocabulary development studies and the need to employ a multi-task approach analysis. The

analysis enables me to report three main findings. First, productive vocabulary development appears to depend to some extent on the specific productive vocabulary task used. Over the three time points, I observed significant differences in overall task scores for the three productive vocabulary task measures (Lex30, G_Lex, and the PVLT). Second, I conducted a detailed analysis of performance within the three frequency bands (2k, 3k, and 5k) shared by the three productive vocabulary tasks. In general terms, I observed an overall improvement in participants' productive vocabulary knowledge. However, I observed decreases for specific tasks and frequency bands (i.e., Lex30 3k, 5k; G_Lex 5k; and PVLT 2k, 3k, and 5k). I suggest this finding relates to the dynamic nature of productive vocabulary growth and potentially alludes to the multiple aspects of knowledge elicited by the three tasks. Third, my comparisons between the different task scores at the different time points, and according to the receptive task (XY_Lex) scores, appears to question the assumption of a consistently strong relationship between receptive and productive vocabulary (e.g., Meara & Fitzpatrick, 2000; Webb, 2008). The evidence from the study suggests that the relationship depends largely on the specific productive (and receptive) vocabulary measures employed.

These results have potentially useful implications for second language acquisition research. In my discussion, I highlighted research (e.g., Daller et al., 2007; Laufer & Nation, 1995) that appears somewhat equivocal with regard to vocabulary knowledge development. Variation might exist for a number of different reasons, not exclusive to the individual, the multiple aspects involved in vocabulary knowledge (e.g., Nation, 2001), the changing relationship between receptive and vocabulary knowledge, or context. I believe the study adds to these findings by adding "task" to the list of influences on productive (and receptive) vocabulary development. The findings also support suggestions from earlier studies that, for example, development occurs differently within specific frequency levels (Webb & Chang, 2012), different aspects of vocabulary knowledge develop differently according to specific frequency levels (Zhang & Lu, 2013), and that use of the PVLT enables researchers to investigate the "developments [that] occur in the different types of vocabulary over a period of time" (Laufer & Nation, 1999, p. 45).

Study Limitations

Before I conclude, I should acknowledge the several limitations with the current study. The first limitation relates to the aspects of knowledge measured. While I might claim that the different tasks elicit different aspects of knowledge, I only do this with reference to

Fitzpatrick and Clenton's (2017) "Vocabulary Test Capture Model." Their model represents a systematic approach to outlining the quality of a learner's word knowledge. With reference to the model, I can estimate that participant response to Lex30 might exhibit the ability to produce word forms or that in response to G_Lex participant responses might demonstrate knowledge of "semantic appropriateness in context." I have added the PVLT to Fitzpatrick and Clenton's model (Figure 5) in which I suggest that correct responses to this task represent knowledge of "semantic appropriateness and grammatical accuracy in context."

Second, the productive vocabulary tasks I report here are written and scores are expressed in terms of frequency. Bearing in mind the limitations of frequency studies, I suggest that future studies concurrently explore productive vocabulary knowledge development according to other indices (e.g., "contextual diversity" (Crossley et al., 2010); "psycholinguistic indices" (Berger et al., 2017); and in comparisons with L2 learner corpora (Monteiro, Crossley, & Kyle, 2018).

A third limitation relates to the fact that the current study only explored the productive vocabulary knowledge of a participant group with the same L1 (Japanese). I cannot, therefore, extrapolate the findings from the current study to other first language communities on the basis that the differences observed here might be language-dependent. I therefore welcome replications of the current study with different populations in order to test the claims I make and to potentially represent a broader picture for second language productive vocabulary development. I would extend this specific limitation to suggest that replications also explore the extent to which different second language proficiency levels exhibit different trajectories. I touch on this in the current study, but sense that larger studies using more advanced students can provide a more substantive picture to second language vocabulary development research. Nevertheless, I believe the current study represents an essential step in investigating the longitudinal development of productive vocabulary knowledge.

Appendices

Appendix 1 Lex30 (Version C)

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 2 G_Lex (Versions B and C)

Version B

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.

G_Lex (Version C)

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 tried to _____ during his summer vacation.					
When I feel angry, I always go to the_____.					
They think football is_____.					
She wanted to _____ the project.					
My best _____is orange.					
She felt _____ when she received her test score.					
She couldn't _____ the house.					
She should include more ____ in her next report.					
My friends feel ____ about my new car.					
The government _____ the people.					
He was surprised about his _____.					
He thought his parents were _____.					
She wanted to _____ her life.					
She sent _____to her boss.					
We looked ____ before the game.					
He wanted to _____ the message.					
She was happy about her _____.					
They thought the basketball game was____.					
He tried to _____ his teacher.					
She gave her mother _____.					
At the graduation party, the family felt ___.					
She always _____her bag.					
Last night, I had my worst _____.					
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. Copyright 2017 by TESOL International Association.

References

Berger, C., Crossley, S., & Kyle, K. (2017). Using Novel Word Context Measures to Predict Human Ratings of Lexical Proficiency. *Educational Technology & Society, 20* (2), 201–212.

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

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

Crossley, S., Salsbury, T., & McNamara, D. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning, 60* (3), 573–605. doi:https://doi.org/10.1111/j.1467-9922.2010.00568.x

Daller, H., Milton, J., & Traffers-Daller, J. (2007). *Modelling and assessing vocabulary knowledge*. London: Cambridge University Press. doi:https://doi.org/10.1017/CBO9780511667268

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. doi:https://doi.org/10.1017/S0272263117000389

Dóczi, B., & Kormos, J. (2016). *Longitudinal developments in vocabulary knowledge and lexical organization*. Oxford: Oxford: Oxford University Press. doi:https://doi.org/10.1093/acprof:oso/9780190210274.001.0001

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

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). London: Cambridge University Press. doi:10.1017/cbo9780511667268.009.

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

Gilabert, R., Baron, J., & Llanes, A. (2009). Manipulating Cognitive Complexity across Task Types and Its Impact on Learners' Interaction during Oral Performance. *International Review of Applied Linguistics in Language Teaching (IRAL), 47*, 367–395. doi:https://doi.org/10.1515/iral.2009.016

Gyllstad, H., & Wolter, B. (2016). Collocational processing in light of the phraseological continuum model : Does semantic transparency matter? *Language Learning, 66* (2), 296–323. doi:https://

doi.org/10.1111/lang.12143

Henriksen, B., & Danelund, L. (2015). Studies of Danish L2 learners' vocabulary knowledge and the lexical richness of their written production in English. *Lexical issues in L2 writing. ed. / Päivi Pietilä; Katalin Doró; Renata Pípalová. Newcastle upon Tyne: Cambridge Scholars Publishing, 2015.*, 29–56.

Housen, A., Bulté, B., Pierrard, M., & Van Daele, S. (2008). Brussels, Investigating lexical proficiency development over time - The case of Dutch-speaking learners of French in. *Journal of French Language Studies, 18*, 1–22.

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

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

Kremmel, B. (2016). Word Families and Frequency Bands in Vocabulary Tests: Challenging Conventions. *TESOL Quarterly, 50* (4), 976–987. doi:https://doi.org/10.1002/tesq.329

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

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

Laufer, B., & Nation, P. (1999). A vocabulary -size test of controlled productive ability. *Language Testing, 16* (1), 33–51. doi: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. doi:10.1111/0023–8333.00046

Meara, P., & Alcoy, O. (2010). Words as Species: An Alternative Approach to Estimating Productive Vocabulary Size. *Reading in a Foreign Language, 22* (1), 222–236.

Meara, P., & Fitzpatrick, T. (2000). Lex30: An improved method of assessing productive vocabulary in an L2. *System, 28* (1), 19–30. doi:10.1016/s0346–

Meara, P., & Miralpeix, I. (2016). *Tools for Researching Vocabulary*. Bristol: Multilingual Matters.

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

Monteiro, K., Crossley, S., & Kyle, K. (2018). In Search of New Benchmarks: Using L2 Lexical Frequency and Contextual Diversity Indices to Assess Second Language Writing. *Applied Linguistics, 41* (2), 280–300. doi:https://doi.org/10.1093/applin/amy056

Nation, P. (1983). Learning vocabulary. *New Zealand Language Teacher 9, 1*, 10–11.

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

Paribakht, T., & Wesche, M. (1993). M. Reading comprehension and second language development in

a comprehension-based ESL program. *TESL Canada Journal*, 11 (1), 9–29. Doi: 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* (pp. 174–200). London: Cambridge University Press. doi: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. doi:https://doi.org/10.1017/s026144481800037x

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

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

Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: A Longitudinal study. *Language Learning, 48* (2), 281–317. doi:https://doi.org/10.1111/1467–9922.00042

Schmitt, N. (2010). *Researching Vocabulary*. London: Palgrave Macmillan.

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

Vasylets, O., Gilabert , R., & Manchón, R. (2017). The Effects of Mode and Task Complexity on Second Language Production. *Language Learning, 67* (2), 94–430. doi:https://doi.org/10.1111/lang.12228

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

Webb, S. (2008). Receptive and Productive Vocabulary Sizes of L2 Learners. *Studies in Second Language Acquisition, 30* (1), 79–95. doi:https://doi.org/10.1017/S0272263108080042

Webb, S. (2009). The Effects of Receptive and Productive Learning of Word Pairs on Vocabulary Knowledge. *RELC Journal, 40* (3), 360–376. doi:https://doi.org/10.1177/0033688209343854

Webb, S. A., & Chang, A. C.-S. (2012). Second Language Vocabulary Growth. *RELC Journal, 43* (1), 113–126. doi:https://doi.org/10.1177/0033688212439367

Zhang, X., & Lu, X. (2013). A longitudinal study of receptive vocabulary breadth knowledge growth and vocabulary fluency development. *Applied Linguistics, 35*, 283–304. doi:https://doi.org/10.1093/applin/amt014

Zipf, G. (1936). *The Psychobiology of Language*. London: Routledge.

Investigating Productive Vocabulary Knowledge Development: A Task-Based Approach

Hosam ELMETAHER

This paper investigates the longitudinal development of productive vocabulary knowledge. I first build on an earlier model to conceptualize productive vocabulary knowledge, adding a commonly used productive vocabulary measure to the model. Second, I report productive vocabulary performance development from the measures examined in the first experiment. I respond to research calling for a development study that concurrently uses measures without inherent problems, is based on research representative of the measure, and employs an analysis of infrequent items. In a multi-task approach, I report that: (i) productive vocabulary development varies according to task and time; (ii) development is inconsistent within three (2k, 3k, and 5k) frequency bands for three productive vocabulary tasks; and, (iii) performance on the tasks varies according to second language proficiency. I discuss the dynamic nature of productive vocabulary and conclude by highlighting implications as well as several potential future approaches for vocabulary development research.