Making sense of learner performance on tests of productive vocabulary knowledge

Tess Fitzpatrick, Cardiff University

Jon Clenton, University of Hiroshima

Making sense of learner performance on tests of productive vocabulary knowledge

Abstract

This paper offers a solution to a significant problem for teachers and researchers of language learning that confounds their interpretations and expectations of test data: the apparent simplicity of tests of vocabulary knowledge masks the complexity of the constructs they claim to measure. We first scrutinise task elements in two widely cited productive vocabulary measures, Lex30 (Meara and Fitzpatrick, 2000) and the Lexical Frequency Profile (LFP, Laufer and Nation, 1995), in order to gain a more precise understanding of the relationship between test performance and learner knowledge. Next, in three empirical studies (N = 80, 80, 100) we compare L2 learners' performance on Lex30, as the static point of reference, with LFP and with two new tests designed to investigate specific elements of the vocabulary test tasks. Correlation analyses indicate systematic differences in the tests' capacity to capture information about the quality of learners' word knowledge and the size of their vocabulary resource. Using the findings from this empirical work, we formulate a model of vocabulary 'capture' onto which test tasks can be mapped. We demonstrate how capturing key elements of the relationship between test scores and lexical competence can guide teachers and researchers in applying and interpreting vocabulary tests.

The apparent simplicity of vocabulary knowledge scores ("this learner 'knows' *n* number of English words") makes them attractive to practitioners needing a quick, efficient way to assess learners' proficiency, progress, or needs. In reality, though, vocabulary test scores represent complex sets of information, and in order for them to be meaningful, subtle and informed interpretation is required. In this paper our objective is to expose and explore tensions between the competences underlying test performance and the demonstration of those competences in specific tests (in this case, tests of 'productive vocabulary knowledge'). Doing this enables us to set out a road map for the interpretation of test scores, which can

support both teachers and researchers in calibrating test scores with other aspects of learner assessment.

Research into second language acquisition does not always map straightforwardly onto teaching practices, but language testing is an area where research outputs are often directly applied in the classroom. The 1980s and 90s saw a flurry of publications presenting tests which had immediate relevance to teaching practice, and which built on earlier work by researchers such as Cronbach (1942), Richards (1976), and Anderson and Freebody (1981). Among the most enduring and influential of these tests are the Vocabulary Levels Test (VLT, Nation, 1983), the Eurocentres Vocabulary Size Test (EVST, Meara & Jones, 1987) the Vocabulary Knowledge Scale (VKS, Paribakht & Wesche 1993), and the Productive Vocabulary Levels Test (PVLT, Laufer & Nation 1999). With the exception of the VKS (which targets bespoke vocabulary items), all of these tests take the same basic approach to capturing vocabulary knowledge, namely they exploit its relationship with frequency. In essence, they assume that learners acquire words according to the frequency with which those words occur in language use. Created on the cusp of the corpus linguistics 'revolution', these tests depend partly or wholly on pedagogical lists such as the General Service List (West, 1953). Twenty-first century iterations of these test types, such as the Vocabulary Size Test (Nation & Beglar, 2007) and version options of the vocabulary profile tools on the Lextutor site (Cobb, http://www.lextutor.ca/) are entirely corpus driven, and are available and easily administered online.

This is, unexpectedly, problematic, because the availability and apparent simplicity of these instruments belies the complexity of the construct they claim to measure. This is increasingly acknowledged in the research literature; influential books by Read (2000) and Nation (2001), and journal articles focusing on specific aspects of vocabulary knowledge (e.g. Laufer, Elder, Hill, & Congdon, 2004; Meara & Wolter, 2004; Webb, 2005; 2007; 2009) have attempted to address this complexity by presenting a more fine-grained conceptualisation of the construct. They share an acknowledgement that knowing a word is not an 'all-or-nothing' phenomenon (e.g. Laufer, 1998) and that vocabulary knowledge must in some manner be seen as multi-dimensional. Within knowledge of an individual word are potentially contained many features, including its definition, collocations, phonological and orthographic representation, affixes, and so on (see Nation's taxonomy of word knowledge, 2001, p. 27). One clear and well-recognised dimension is the trajectory from receptive to

productive knowledge (e.g. Melka, 1982). Another is the dynamic progression from partial to precise word knowledge, and an awareness of how a word relates to others in the lexicon (Henriksen, 1999). Finally, as a product of these, we must recognise the way that vocabulary knowledge does not only concern individual words, but also includes the way they are organised in the mental lexicon (Meara, 1996) and, related to this, speed and, ultimately, automaticity of retrieval (Qian, 2002). Each conceptual dynamic entails challenges for the development and interpretation of tests, and this paper engages with that debate with respect to the notion of productive vocabulary knowledge.

The distinction between productive knowledge (sometimes referred to as active knowledge, or recall) and receptive knowledge (passive knowledge, or recognition) is one of the most pervasive subdivisions in vocabulary knowledge research. The labels map onto the ubiquitous contrast in the communicative language teaching community between reading and listening (receptive) and speaking and writing (productive) skills, and in that sense are familiar to researchers and practitioners alike. They are relatively uncontended labels, and the nature of classroom (as opposed to laboratory) vocabulary testing fits the receptive/productive distinction comfortably, with any given test eliciting vocabulary knowledge through one of the four skills. At its most simplistic, this means that in tests, learners are required either to demonstrate their understanding of a given item (which they have heard or read), or to produce the item (by saying or writing it) in response to a cue of some kind. Vocabulary tests which identify themselves as targeting receptive knowledge include the Vocabulary Levels Test (VLT, Nation, 1983), the Eurocentres Vocabulary Size Test (EVST, Meara & Jones, 1987), X_Lex (Meara & Milton, 2003), AuralLex (Milton & Hopkins, 2006), and the Vocabulary Size Test (Nation & Beglar, 2007). Those which use the 'productive' label include the Lexical Frequency Profile (LFP, Laufer & Nation, 1995; 1999), the Productive Levels Test (PLT, Laufer & Nation, 1999), Lex30 (Meara & Fitzpatrick, 2000), and P_Lex (Meara & Bell, 2001). Within the 'productive' category, researchers have attempted to make more informative distinctions by using the subcategories 'controlled productive' and 'free productive' to identify the specific aspect of knowledge being targeted (Laufer, 1998; Laufer & Nation, 1995; 1999; Laufer & Paribakht, 1998; Nation, 2001). 'Controlled' indicates that the test is designed to elicit specific, predetermined, vocabulary items, and 'free' indicates that vocabulary produced by the test-taker in a relatively unconstrained task will be measured.

MAKING SENSE OF VOCABULARY TEST PERFORMANCE

Most of the receptive vocabulary tests cited above assess vocabulary "size": the number of words a learner knows, at threshold level at least. They typically do this by testing knowledge of a sample of words designed to represent frequency bands in a principled way, and by using formulae to extrapolate overall vocabulary size from this. Productive vocabulary tests, though, are less straightforward to interpret. If a principled sample of target words is preselected (as in 'controlled' tests), test prompts must give enough information to elicit the target, but not so much that production of the target item is scaffolded or assisted, and if inferences are to be drawn about un-tested words, many test items will be needed. If the test is of 'free' productive knowledge, claims that a representative vocabulary sample has been produced are difficult to support (see Meara and Olmos Alcoy (2010), who describe how repeated sampling might be used to extrapolate more realistic estimates of overall knowledge, by adapting Petersen's ecological 'capture-recapture' method). Because representativeness and sample size underpin extrapolation to claims about the lexicon in general, design and interpretation of these tests is challenging. Nation and Webb (2011) recognise this, suggesting that tests of productive vocabulary knowledge might be more problematic than their receptive equivalents because (a) they tend not to give credit for partial word knowledge, and are therefore less sensitive (2011, p. 304), and (b) they are unlikely to relate easily to vocabulary size, because size estimates depend on the production of a meaningful sample of words of different frequencies, and calculations can be confounded by text length and genre (2011, p. 200-201). Nevertheless, it is important to note that there are aspects of language knowledge and performance which productive vocabulary tests are well-suited to tap into, including those represented in taxonomies such as Nation's (2001, p. 27, see above), and considerations relating to fluency, lexical availability, and the developmental relationship between receptive and productive knowledge.

In the light of these issues, this paper has both a substantive and a theoretical aim. The substantive aim is to identify precisely what is being measured by tests that claim to target 'productive vocabulary knowledge'. To do this, we begin by comparing learner performance on two widely-cited tests of productive vocabulary knowledge: the Lexical Frequency Profile (LFP), created by Laufer and Nation (1995), and Lex30, created by Meara and Fitzpatrick (2000). We describe and compare the two tests, and then present a study comparing learner performance on them. Questions raised by the findings of that study are addressed in two further empirical investigations. The theoretical aim is to reveal accurately and efficiently the

capacity of different tests to capture the quality of learners' knowledge of individual words, and the size of their vocabulary resource as a whole. By addressing both these aims in a single account, we can benefit from a more precise understanding of the relationship between test performance and learner knowledge, and use this to formulate a model that captures key components of this relationship.

Study 1: Comparing Learner Performance on Lex30 and the Lexical Frequency Profile Tasks.

In Study 1 we compare learner performance on two tests relating to productive vocabulary knowledge. We begin this section with a description and a comparative analysis of the tests, and then present a study in which both tests were administered to a single group of learners.

The Test Tools

Lex30. The Lex30 test uses a word association format, presenting learners with a list of 30 stimulus words in English, and instructing them to "write down the first four (English) words you think of when you read each word in the list". The test was created by Meara and Fitzpatrick (2000) as a means of estimating the productive vocabulary knowledge of English language learners, and has been trialled and evaluated in subsequent studies, including Fitzpatrick and Meara (2004); Fitzpatrick (2007); Fitzpatrick and Clenton (2010); Jiménez Catalán and Moreno Espinosa (2005); Walters (2012).

Meara and Fitzpatrick used criteria for cue selection that maximised the likelihood of eliciting a high proportion of varied and infrequent responses (2000, p. 22). Specifically, (a) cues were taken from the first 1000 most frequent English words, minimising the risk of learners encountering words they do not know; (b) words that tend to elicit the same response from everyone were excluded, so as to maximise the opportunity for differentiation between test takers; and (c) words which typically produced high-frequency responses were not eligible as cues, in order to allow subjects as much opportunity as possible to produce infrequent vocabulary items. Response data from the Edinburgh Associative Thesaurus (a set of word association norms compiled by Kiss, Armstrong, Milroy, and Piper J., 1973) were scrutinised to ensure cues met criteria (a) and (b) (see Meara and Fitzpatrick (2000) for a detailed account of cue selection).

The responses, which amount to up to 120 per test taker (30 x 4), are categorised according to frequency. A mark is given for each "infrequent" vocabulary item produced, with infrequent being defined as "not in the first 1000 most frequent English words". The final score can be expressed as a tally (out of a maximum of 120) or as a percentage (of all words the learner has produced, which may be fewer than 120). See Appendix A (in Supporting Information) for an example of a completed Lex30 task.

The Lexical Frequency Profile (LFP). The LFP measure also categorises learner output according to frequency, but in this test the learner's output is generated through an essay response to a discussion question. Laufer and Nation (1995) designed the test as a measure of vocabulary use, and required their subjects to write two compositions, of 300-350 words, in successive class periods. The first composition question was: *'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*. For the second question subjects could choose from three further topics (see Appendix B in Supporting Information).

The compositions were processed according to four criteria: (a) if a word was clearly used incorrectly, it was excluded from analysis; (b) misspellings were corrected; (c) incorrect derivatives were tolerated as examples from their word family; and (d) proper nouns were excluded from analysis. The first 300 words of each composition were then categorised as belonging to one of four frequency-related bands: the first 1000 most frequent words (1k), the second 1000 (2k), the University Word List, and 'not in lists', thus creating, for each composition, a 'Lexical Frequency Profile'. To facilitate statistical comparison with other single-score tests, Laufer and Paribakht suggest the use of a 'condensed profile', representing the percentage of beyond-2000 words, or "the sum of the percentages from the University Word List (UWL) and "not on the list"" (1998, pp. 374-375). Because the study we present here entails a comparison with a single score test, this is the approach we have adopted.

Scoring protocols. Since the purpose of our study was to compare the profiles of learners across different tests, we used an identical scoring protocol for both tests. The task data were entered into a computer text file and submitted to the WebVP at www.lextutor.ca (Cobb, n.d.), to be categorised according to the number of items within each of four frequency levels: the first thousand most frequent word families; the second thousand word list; the academic word list (AWL); and words that do not appear on the other lists (Off-list words). The WebVP was used because it categorises in line with the word lists used by

Laufer and Nation (1995) in scoring the Lexical Frequency Profile. LFP scores were generated according to Laufer and Paribakht's 'condensed profile' calculation (1998), by counting items produced outside of the 1k and 2k bands. The Lex30 scores were generated, also using WebVP, by counting items produced outside of the 1k band. (This inconsistent definition of "infrequent" words is addressed systematically in the studies below). As subjects vary in terms of how many words they produce, and to minimise the effect of that variable, we use percentage scores for all analyses.

LFP and Lex30: a comparison of test formats. The LFP assesses vocabulary produced in an essay-writing task, and Lex30 assesses vocabulary produced in a word association task. Both tests have been available to the research community for a long enough period of time to have been subjected to scrutiny by a range of researchers, in a number of study contexts. Until now, no single study has compared learner performance on these two tests, but there are reasons why one might hypothesise that learners' scores on LFP and Lex30 will be mutually predictive. Read's framework for test comparison (2000, pp. 7-13) conceptualises vocabulary assessment in three "dimensions": knowledge construct, item selection and context dependence, and application of this framework offers an appropriate starting point for our comparison. In terms of item selection, both LFP and Lex30 are "comprehensive measures" in that they do not focus on specific vocabulary items; that is, the tester does not have a predetermined list of items that the testee must produce. Regarding knowledge construct, both tests represent a "discrete" approach to vocabulary knowledge, measuring it as an "independent construct" (rather than it being embedded within an assessment of, for example, reading or writing). The role of context (Read's third dimension) in these tests is less easy to identify: the Lex30 cues, the LFP essay title and the LFP running text can all be interpreted as context, but test-takers' degree of engagement with the context is difficult to ascertain, and likely varies between participants and items. An additional shared feature to note, though, is that both tests use the same external referent to calculate performance: the frequency of occurrence of vocabulary items in general usage. They are both, therefore, underpinned by a model of vocabulary acquisition whereby the order in which learners acquire words aligns with the frequency of those words in language use (the findings of, for example, Aizawa (2006) and Aizawa and Iso (2007) support this model). Finally, there is existing empirical evidence from a wide range of studies that each of these tests not only passes validity standards in its own right (Fitzpatrick & Meara, 2004;

Fitzpatrick, 2007; Fitzpatrick & Clenton, 2010; Laufer & Nation 1995; Laufer & Nation 1999; Laufer & Paribakht 1998; Walters, 2012), but also correlates positively and significantly with scores from a third test of productive vocabulary, the "Productive Vocabulary Levels Test' (PVLT, Laufer & Nation, 1999). The LFP-PVLT correlation is reported in Laufer and Paribakht (1998), and the Lex30-PVLT in Fitzpatrick (2007).

Despite these similarities, from the perspective of the test taker LFP and Lex30 require the completion of very different tasks. For LFP s/he must write a 300-word discursive essay. Lex30 is a word association task, requiring four single lexical items to be written in response to each of 30 cue words.

It is clear, then, that a tension exists between the tests' shared characteristics (free productive vocabulary knowledge, a discrete approach, use of frequency measures) and the differences in the elicitation techniques they use (discursive for LFP, single item for Lex30). It can be argued that these evident differences are in fact inconsequential, since both tests elicit data in different ways but then treat it in the same way. But that argument rests on the shaky assumption that the same data will be generated irrespective of the task, or more conceptually, that a test taker's vocabulary knowledge is tapped in a manner independent of the mode of elicitation.

For this reason, the tension between what is shared and what is not shared across the tasks is worthy of systematic examination, and below we report how we undertook that investigation. We took an incremental approach using three consecutive studies, the first of which compares learners' Lex30 and LFP scores. The research questions addressed in studies two and three are each generated from the findings of the previous study, and by using a suite of studies we are able to separate out constituent elements of the tests for closer scrutiny.

Is there a Correlation between the Scores in Lex30 and LFP? Study One Procedure and Results

The participants in Study 1 were 80 (26 female, 54 male) L1 Japanese learners of English, aged between 18 and 21. They were university students in Medicine and Engineering faculties, and received three hours of English classes per week. Their English language proficiency was rated by their teachers as pre-intermediate to intermediate; the learners had received approximately three hours of English language tuition weekly from the age of 13, and scores on an independent TOEFL test were in the 420-480 range. Learners completed the Lex30 and the LFP task within two class periods. In the first they took the

Lex30 task and first LFP composition task, and one week later they took the second LFP composition task. The test data were scored according to the protocols described above, so that each score represents the percentage of "infrequent" words produced. (Note that the condensed LFP profile defines 'infrequent' as outside the 1k and 2k bands, whereas Lex30 defines it as outside the 1k band. Although on the face of it these different interpretations of 'infrequent' seem to introduce an inconsistency, we retain them in the initial analysis below in order to maintain the integrity of the individual tests).

Table 1: *Lex30 and LFP scores (N=80)*

	Mean score	SD	Minimum score	Maximum score
Lex30 (%)	43.63	5.89	29.41	57.83
LFP (%)	5.28	2.3	0.69	11.22

Learner performance on the two tests is shown in Table 1. The correlation between Lex30 and LFP scores is not significant (r = 0.186, p = .098). As noted above, there is an inconsistency in the way the two tests define 'infrequent'. To investigate whether this discrepancy impacted on the relationship between test scores, we recalculated the LFP scores to award points for items produced outside 1k (thus defining 'infrequent' in the same way as Lex30, and accommodating Laufer and Nation's note that for less proficient learners a meaningful distinction can be made between words produced in the first and second thousand bands (1995, p. 311)). Despite this adjustment and the consequent increase in mean LFP scores (see Table 2), the correlation between Lex30 and the recalculated LFP scores remains non-significant (r = 0.108, p =.339).

Table 2: *LFP adjusted scores (with infrequent defined as beyond 1k)*

	Mean score	SD	Minimum score	Maximum score
LFP >1k (%)	10	3.14	3.61	18.68

It seems, then, that the differences in elicitation techniques used in the two tests mean that the vocabulary samples yielded by the LFP and Lex30 tasks do not represent the learner's lexicon in equivalent ways. As noted above, the LFP task is a discursive one and as such inevitably elicits function words (most of which are "frequent"), and this influences the proportion of infrequent items participants produce. Lex30, on the other hand, instructs participants to write single word responses, and elicits almost no function words (none at all from most participants). Furthermore, a subject's opportunity to produce infrequent vocabulary items in the LFP task might be limited by other perceived discursive considerations, such as the demands of register, topic, and cohesion (see Leech, 1994 on the complex factors affecting word choice in learner essays).

In order to investigate the degree to which these discursive considerations contributed to the findings of Study 1, we created a task that follows the elicitation cue and the scoring protocols of the LFP, but that does not require the learner to produce a discursive text. In the following section we present this task, the "Brainstorm Frequency Profile (BFP)", and report Study 2, which compares learner performance on BFP and Lex30.

Study 2: Comparing Lex30 with a Modified Version of the Lexical Frequency Profile, Designed to Elicit Vocabulary in a Non-Discursive Way

Our second study explores the hypothesis that the LFP scores do not relate straightforwardly to Lex30 scores in Study 1 because of the constraints on word choice imposed by the discursive considerations of essay writing. For Study 2 we designed a non-discursive equivalent to the LFP task (i.e. one in which words are elicited with no reference to syntactic context): the "Brainstorm Frequency Profile (BFP)".

Test Tool: The Brainstorm Frequency Profile (BFP)

The BFP uses the same topics as the LFP, but instead of essays, test takers are asked to write down as many single words, relevant to the topic, as they can; whereas the original LFP question cue is "Discuss this idea", the BFP task instruction is "Write as many one-word responses as possible to this idea". By removing the need to construct a coherent text, the BFP more directly accesses the test taker's vocabulary knowledge, since there is no competition for attention from grammatical or discursive requirements. An example response from the BFP can be seen in Appendix C (Supporting Information). The words given in responses were profiled using the same procedure as the LFP.

Is there a Correlation between the Scores in Lex30 and the Brainstorm Frequency Profile? Study Two Procedure and Results

A new group of 80 (8 female, 72 male) L1 Japanese learners of English participated in Study 2. They were selected based on their similarity in profile to Study 1 participants, and had the same language learning background and teacher-rated proficiency (TOEFL scores were in the 410-470 range). All were university students in Technology or Engineering faculties. The participants completed the Lex30 and then the BFP in two class periods, with a one-week gap between test times. As in Study 1, task scores represent the percentage of "infrequent" words produced; for Lex30, "infrequent" was defined as outside the first 1000 most frequent words. BFP was scored first in line with the LFP scoring protocols (with "infrequent" defined as outside the 2k band), and then, as in Study 1, rescored defining "infrequent" as outside the 1k band.

Table 3: *Lex30 and Brainstorm Frequency Profile scores (N=80)*

	Mean score	SD	Minimum score	Maximum score
Lex30 (%)	38.5	5.9	23.6	47.6
BFP >2k (%)	22.1	8.5	0.0	39.7

As can be seen in Table 3, the two tasks produced different score profiles, with the BFP eliciting a smaller average percentage of infrequent words, and a greater range of scores. The correlation between the two sets of test scores is not significant, (r =0.153, p =.175). Again, a possible explanation for the differences in performance is that different frequency bands are used to score the two tasks; we rescored the BFP task data in accordance with the same frequency bands used to score the Lex30 task (one mark for every word outside the 1k band). The reanalysis yielded an average BFP score of 28.1% with the correlation between the two sets of scores remaining non-significant (r = 0.211 p =.061), indicating that the differences between performance on the two tests cannot be explained by the difference in original scoring systems. This finding suggests that the differences in learner performance on Lex30 and LFP in Study 1 is not accounted for by the requirement by the latter to produce vocabulary in sentence context.

In Study 3 we investigate the hypothesis that the difference in test performance identified in Study 1 is essentially one of sampling, and relates to systematic differences in the quantity of elicitation prompts used.

Study 3: Comparing Lex30 with a Task in which Words are Elicited using Contextual Priming and Multiple Prompts

In Study 3 we design a tool which, like LFP, requires test takers to attend to context (both semantic and syntactic) in completion of the test task. However, in order to investigate the degree to which sampling method affects test performance, this tool uses multiple elicitation prompts. Theories of lexical activation and access inform us that "conceptual preparation" is a prerequisite for lexical selection and, ultimately, articulation (e.g. Levelt, Roelofs & Meyer 1999, p. 3), and it follows that by varying conceptual activation, a wider range and number of lexical items will be made eligible for selection. In both the BFP and LFP task, there is a single conceptual activation event, in the form of the task question, and it is reasonable to hypothesise that the pool of lexical items eligible for selection from this prompt is exhausted during the task. The 30 one-word cues of Lex30, on the other hand, constitute 30 activation events, with a new set of lexical items eligible for selection in each new event.

This interpretation drives the design of our third study, in which we pilot 'G_Lex', a gap-fill vocabulary test which, like Lex30 but unlike the LFP or BFP, has multiple 'activation events'. G_Lex, like the LFP, matches Lex30 in Read's first two dimensions (i.e. it is discrete and comprehensive). G_Lex is unlike Lex30, though, in terms of Read's third dimension: G_Lex requires test takers to attend to context (both semantic and syntactic) in completion of the test task. The production of words in the LFP task, as suggested above, is similarly constrained; in LFP, though, the contextual constraints are imposed by the learner him/herself as s/he composes the text. In G_Lex the context is provided in the test tool. This is advantageous in two ways: (a) the test taker is not distracted by attempts to mediate the constraints of context (e.g. "I won't use that structure because I can't remember which verb form follows it", or "I won't use that word because I don't know whether it collocates with x or y), and (b) the context is consistent across all learners, because it is embedded in the test tool.

The Test Tool: G_Lex Gapfill Task

In the G-Lex task, test takers are given 24 sentences, each containing one gap, and are required to provide up to five different words that might fit into each gap. The gaps are suitable for nouns, adjectives and verbs in equal measure (8 sentences each). The twenty-four sentences are formulated in order to minimise receptive processing load, and to maximise the test's capacity to distinguish effectively between learners with different degrees of (productive) lexical resource. Specifically the sentences meet the following criteria: (a) they are syntactically simple; (b) they contain only high frequency words; (c) they readily elicited five responses from native speakers or proficient non-native speakers in a pilot test; (d) they did not elicit lexical sets (e.g. *banana, apple, orange,* etc.); (e) they did not elicit similar words to another sentence in the task. Both native and non-native speaker groups piloted the G_Lex sentences, enabling rejection of sentences that did not meet these criteria. We also rejected sentences if they elicited too few responses or only highly frequent responses. Figure 1 shows the task instruction and the first five test items. A completed version of the G_Lex task is in Appendix D (in Supporting Information).

'In the spaces provided below write as many one-word responses as possible (up to five) to complete each sentence. Try not to repeat words you have already used.'

1. She loved to _____ over the phone.					
2. When I feel sad I always go to the_____.					
3. They think car-racing is_____.					
4. His colleague wanted to _____ the report.					
5. My favourite _____ is football.					

Figure 1. G_Lex task instruction and first five test items

G_Lex aims to elicit the same potential maximum number of responses as the Lex30 task (120), and participants were therefore given the same amount of time, 15 minutes, as for Lex30. Scoring of the G_Lex task responses also followed the same protocol as that established for Lex30. In other words, responses are accepted if they are spelled accurately

enough to be identified, and the score is calculated according to the number of infrequent (>1k) words produced.

Is there a Correlation between the Scores in Lex30 and the G_Lex Gapfill Task? Study Three Procedure and Results

One hundred L1 Japanese learners of English (43 female, 57 male), identical in age, language learning background and proficiency level as those in the studies above (though this time from a wider range of University faculties), completed the Lex30 and the G_Lex tasks within two class periods, with a one-week interval between test times. Test output was scored in the same way as Studies 1 and 2, with one point awarded for every item outside the 1k band. This tally was then converted to a percentage of all words produced.

Table 4 - *Lex30 and G_Lex task scores (N=100)*

	Mean score	SD	Min score	Max score
Lex30 (%)	32.5	8.8	13.1	55.0
G_Lex (%)	29.1	8.9	10.7	53.0

The mean scores and standard deviations, shown in Table 4, indicate that similar proportions of infrequent vocabulary items are produced in response to the Lex30 and G_Lex tasks. Scores on the two tasks correlate significantly at $r = 0.645$ ($p<.01$), indicating that performance on one task is moderately predictive of performance on the other. Below we discuss theoretical implications of this finding, and of the findings from the first two studies.

Interpreting Score Data from Tests of Productive Vocabulary Knowledge

At the start of this paper we identified an important challenge for the measurement of productive vocabulary knowledge, namely whether the different tests of it that are based on frequency profiling are measuring the same thing. We acknowledged, but easily transcended, superficial differences such as which responses count as ‘infrequent’, and moved into the examination of two main features: discursive vs. non-discursive test tasks, and sampling, or the number of different ‘dips’ into knowledge that are stimulated during a test. We found that the frequency scores from responses in Lex30, which are decontextualized but activate a variety of semantic areas, differ from those generated by the same learners using the LFP and

also the BFP tests, but are compatible with those from the G_Lex. In the light of these findings we discuss, below, three matters that are crucial to the effective design and deployment of vocabulary tests: the advantages and constraints of using a frequency paradigm, the relationship between test design and elicitation, or 'capture', of vocabulary items, and the fitness for purpose of the term "productive vocabulary knowledge".

Test Design and the Frequency Conundrum

We open this discussion by focusing on the constraints that use of a frequency paradigm impose on test design. Lex30 and the LFP derive scores from the number of infrequent words produced by test takers in response to the test task, as do the additional tasks we designed for the investigations reported above (studies 2 and 3: the BFP and G_Lex). This accords with a frequency-driven order of acquisition model, and is inferential in nature: a learner with mastery of, say, the 6000 word frequency band is assumed to have an equal or better degree of mastery of the 1000-5000 bands. The tests used in our studies class all items beyond the first thousand band as 'infrequent', and the greater the proportion of infrequent items produced in response to a prompt, the more words it is assumed the test taker knows. Because the items produced are taken to represent the words the test taker knows, the sampling process is key. However, the lack of correlation between Lex30 and LFP scores, and between Lex30 and BFP scores, indicates that those tests do not tap into the same qualities of word knowledge, and therefore do not sample the learner lexicon, in the same way.

When we consider the mean proportion of 'infrequent' words produced by learners in each of our studies (Table 5), we see that the percentage of infrequent words elicited by the LFP is noticeably smaller than by the other tests.

Table 5 - *Proportion of infrequent (>1k) words produced*

	Lex30	LFP	BFP	G_Lex
study 1	44%	10%		
study 2	38%		28%	
study 3	33%			29%

This can be explained by the discursive nature of the LFP task; syntactic structures require the use of function words, which are typically high-frequency. While this confounds

comparison with single item elicitation tasks we note that this characteristic of the LFP is also a strength: in measuring the 'performance' of vocabulary knowledge in a realistic task (which Laufer (1998) likens to vocabulary use "in letters, reports, oral presentations" (p. 258)) the test is afforded ecological validity.

The relatively low scores yielded by the LFP task are nevertheless problematic: in all the tests used in the studies above, the majority of items produced will be high frequency (within the 1000 band). But in order to gain a sense of the extent of learners' lexical resource beyond this band, opportunity for them to produce infrequent items should be maximised. As we noted earlier, this is a challenge particular to the testing of "free productive vocabulary": in tests of receptive knowledge items can be selected to represent different frequency bands, as they can in "controlled productive tests" such as the Productive Vocabulary Levels Test, (Laufer & Nation 1999). Use of the frequency paradigm in vocabulary testing is constraining in a further sense: lexical improvement can also be represented in extended uses of (frequent) words, morphological sophistication, collocational awareness, and so on, none of which are captured in the kinds of frequency profiles we are investigating in this paper. (Horst and Collins 2006 using LFP in a longitudinal study report that while the proportion of words produced in the 1000 band does not decrease over time, learners' lexical production does become more "register appropriate, diverse, and morphologically complex" (p. 102)).

This dual development trajectory, of the quality of individual word knowledge on the one hand and of the quantity of words known, on the other, goes to the heart of our objective in this paper: to identify what the creators and users of these tests are actually capturing under the label of productive vocabulary knowledge. The frequency paradigm addresses the quantity of words known, but the quality of individual word knowledge influences the likelihood of a learner actually producing an item in response to a particular prompt. In the following section we develop a model for evaluating the scope of individual vocabulary tests against these two dimensions.

Mapping the 'Capture Zones' of Productive Vocabulary Tests

The empirical studies reported in this paper have revealed substantive differences in test function, and highlight the need for teachers and researchers to take an informed approach to the interpretation of test scores. We have explored two dimensions on which the capacity of productive vocabulary knowledge tests to 'capture' vocabulary knowledge might

differ, and these can broadly be conceptualised as the number of words the learner has the capacity to produce, and the quality, depth or thoroughness of (individual) word knowledge. In order to illustrate the second of these, we will adapt a model used to assess learners' knowledge of bespoke word lists: the vocabulary knowledge scale (VKS, Paribakht and Wesche, 1993; 1997, p. 181). Using self-report followed by interview, the VKS ranks learners' knowledge of a word on the following scale:

1. the word is not familiar at all
2. the word is familiar but its meaning is not known
3. a correct synonym or translation is given
4. the word is used with semantic appropriateness in a sentence
5. the word is used with semantic appropriateness and grammatical accuracy in a sentence

The scale is one directional and partially implicational (if a word is known at level 5, knowledge aspects 3 and 4 are assumed), and the frequency paradigm predicts that highly frequent words will achieve level 5 knowledge before less frequent words. The VKS was designed to measure incidental learning of specific vocabulary items through reading, but here we will adapt it to focus on non-specific productive vocabulary knowledge. We do this by (a) replacing the implicational scale with discrete levels of (incomplete) word knowledge, but retaining the directionality (a word with level 5 status will have been at level 4, 3, and so on previously in the acquisition process); (b) relating levels of knowledge to productive word use; (c) adding a quantitative dimension that relates to the number of words known at each level. The resulting model is shown in Figure 2.

Through scrutinising the processes engaged by a test task, and the number of items activated for potential production in the task, we can map specific tests onto the model. The less overall height covered by the test footprint on the vertical axis, the more precise we can be about the kind of knowledge it is measuring. The more broadly it maps onto the horizontal axis, the more confident we can be that knowledge of the specific test items is representative of the whole of that learner's lexicon. Figure 2 illustrates how Lex30 and the LFP might map onto this model.

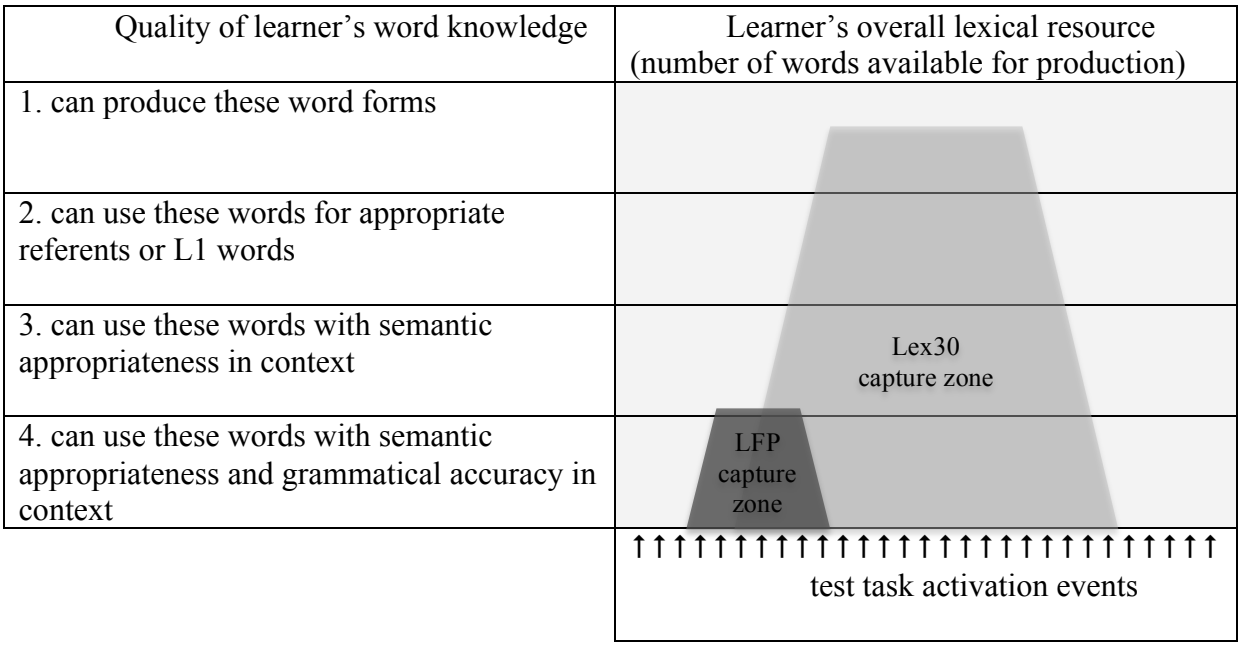

Figure 2. Vocabulary Test Capture: Lex30 and LFP

The vertical dimension of the Vocabulary Test Capture model in Figure 2 relates to the nature of the test task: words are likely to be produced in the discursive LFP task only if the learner feels s/he has semantic and grammatical mastery of their use (Laufer and Nation note that incorrect use of words was rare in LFP data (1995, p. 315)). Lex30 requires single word responses, so learners may use words at level 2 as well as 3 and 4, and may even include level 1 words if they are associated with the cue by a route other than meaning (e.g. form or collocational link). For beginner learners, with vocabularies of less than 1000 words, all levels of the scale are likely to be populated with high frequency words. As the learner lexicon grows, lower frequency words will enter the lexicon and progress through the levels. The capture zone of Lex30 is, therefore, likely to contain a higher proportion of infrequent words than the capture zone of LFP; this is borne out by the findings of our Study 1.

The horizontal dimension of the model relates to the proportion of the lexical resource that the test task has capacity to capture. For levels 1-4 above, we can take lexical resource to be the 'number of words available for production in this learner's lexicon'. The breadth of the test zone reflects the extent to which the sample elicited by the test can be taken as representative of all the items a learner has the capacity to produce at each level, across various functions, contexts and topics. Here we return to the notion of 'activation events': the prompts used to elicit vocabulary in these test tasks. The LFP uses a single prompt, an essay

title, and this limits the words likely to be produced to those related to the essay topic, hence the relatively narrow capture zone in relation to Lex30 which uses 30 prompts, each activating a different semantic field.

We can test the model further by mapping onto it the two other test tasks we created for Studies 2 and 3 above. The BFP requires single word responses to the same activation prompt as the LFP, so it is likely that the horizontal dimension of the zone will be similar to that of LFP. However, no attention is needed to semantic appropriateness or grammatical accuracy, so the vertical dimension of the zone may be more similar to that of Lex30. Unlike Lex30, though, words produced in the task are activated by the ideas the learner has generated in response to the question they have been asked, and, following Flower and Hayes (1981) Cognitive Process Model, are partial "translations" of these ideas (with translation defined as "putting ideas into visible language" p. 373). Responses will not, therefore, include any words at level 1 (where only the word form is known, and is activated by a route other than meaning-based).

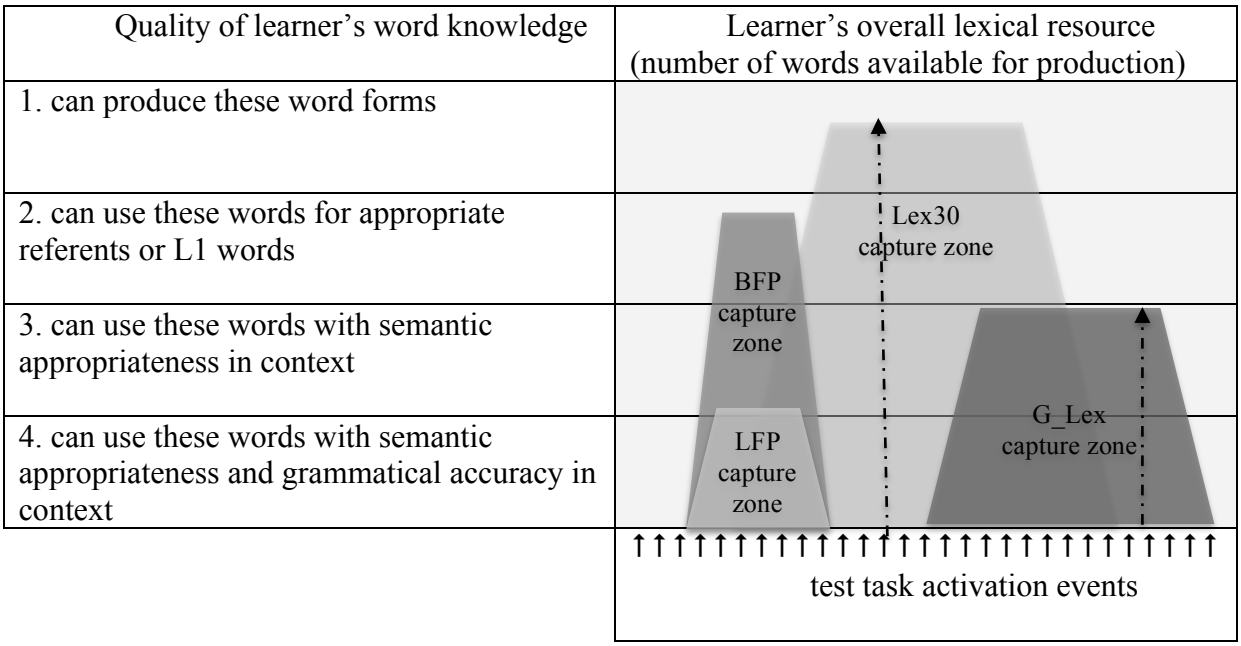

Figure 3. Vocabulary Test Capture: Lex30, LFP, BFP and G_Lex

MAKING SENSE OF VOCABULARY TEST PERFORMANCE

We now consider how learner performance on the G_Lex task, used in our third study, fits the 'capture' model. Recall that G_Lex task was designed to include multiple activation events (24, each relating to a different semantic field), and to elicit nouns, verbs, and adjectives in equal measure. Together this suggests a relatively broad capture zone, with potential activation of many candidate words. However, the G_Lex differs from Lex30 in its requirement for engagement with (sentence) context, and for this reason is unlikely to elicit words at knowledge level 2: test takers are only likely to produce words for which they have semantic and grammatical mastery. Figure 3 suggests how the BFP and G_Lex tests might map onto the capture zone model.

Mapping the capture zone of each test onto a two-dimensional model of knowledge quality and lexical resource size creates a new conceptual space for understanding why we do not find consistent correlations between scores produced by this battery of tests, all of which claim to measure productive vocabulary knowledge. Using the activation events of Lex30 and G_Lex, we can probe the model further: each activation event (cue word, for Lex30 and sentence prompt, for G_Lex) demands multiple responses (4 and 5 respectively) and this requires the learner to dip again and again into the same subset of lexical resource, pulling out consecutive items that are closely related. These 'multiple dip' activation events are represented by the long arrows in Figure 3. We might posit that each time the learner revisits the lexicon, s/he has to probe deeper into zones of less complete knowledge, where infrequent words are more likely to be found. To claim that this, and the dimensional mappings of the tests onto the model in Figure 3, explain the correlation found between Lex30 and G_Lex scores would be to over-extend the interpretation of findings in this paper, but we suggest that the vocabulary test capture model offers a means of conceptualising and exploring the differences and similarities between test tools in a holistic and transparent way.

Determining the optimal way of presenting this conceptual space requires consideration of a number of factors, not yet completely settled but framed, now, in a way that facilitates further investigation. For example,

- Should the positioning of the capture zones on the horizontal axis be, as thus far, arbitrary, or could some calibration be developed?
- How distinct is it possible to make the boundary between knowledge levels 2 and 3 in terms of how we conceptualise vocabulary knowledge, and how we operationalize the model for the purposes of teaching and research?

- Is it safe to assume that vocabulary items are always acquired in the order 1 to 4, and under what circumstances might they not be (e.g. producing internally complex formulaic expressions, not all of whose individual components are known)?
- Does the model actually imply that words are learned in the order 1 to 4, since it is a snapshot of current knowledge, not of how that knowledge was developed? Could modifications to the model make that distinction clear?
- Is the four-level 'quality of knowledge' measure sophisticated enough to account for what is acknowledged to be a multi-faceted and complex notion?

These questions, conceptual and practical, help highlight the potential for further work to develop and modify the model. Yet even in its present iteration, the model offers significant opportunities for development and application, enabling the exploration of features such as

- The role of 'lexical availability' in the model; in particular, the prediction that the most available items in the lexicon are those that are most thoroughly known (i.e. at level 4), can be tested empirically.
- The mapping of frequency onto the model; for beginner and intermediate learners, we predict that infrequent words lag behind frequent ones in their progression through the levels, but test capture might operate differently for proficient learners, who have a high proportion of lexical resource at level 4.
- The mapping of proficiency onto the model; the model currently depicts an equal 'size' of lexical resource at each of the four levels; the distribution of resource is likely to change as proficiency develops, and this can be tested empirically.

Conclusion

We conclude our discussion by considering, again, the construct of productive vocabulary, and the fitness for purpose of the term 'productive vocabulary knowledge'. In the model presented above we have conceptualised the learner lexicon as comprising words at four levels of knowledge, and we suggest that words are available for production at each of those levels, but elicitation of these items will depend on (a) how much contextual engagement is required and (b) how many conceptual activation events occur. Productive vocabulary knowledge in the broadest sense, then, might include all words with capacity to

be activated in some way (not necessarily through knowledge of meaning) that causes the learner to speak or write them. In considering the Lex30 task a test of productive vocabulary, we are applying this broad approach. If we conceptualise productive vocabulary knowledge as entailing knowledge of constraints of and opportunities for word *use*, we need a test such as LFP, that demands close conceptual engagement. Fulcher defines constructs as "the abilities of the learner that we believe underlie their test performance, but which we cannot directly observe" (2010, p. 96), and in the section above we presented a model that supports the mapping of "test performance" (the capture zones) onto the learner's underlying "abilities" relating to vocabulary production (knowledge levels and resource). The empirical and theoretical work we have reported in this paper originated in our challenge to the notion of "productive vocabulary knowledge" as a unitary construct, and demonstrates a means of capturing these different aspects in a conceptual and practical manner. Finally, it reveals the imperative, and provides a means, for teachers and researchers to identify the level and range of the knowledge being tested, if they are to make sense of learner performance on tests of productive vocabulary.

ACKNOWLEDGEMENTS

We would like to thank Alison Wray for her valuable and insightful comments on an early draft of this paper. We are also grateful to the two anonymous reviewers of our paper for their feedback.

THE AUTHORS

Tess Fitzpatrick is Professor of Applied Linguistics at Cardiff University. Her research interests include the investigation of lexical acquisition, attrition, and retrieval processes (with a focus on word association behaviour), the creation and evaluation of vocabulary knowledge assessment tools, and the design and application of innovative language learning techniques.

MAKING SENSE OF VOCABULARY TEST PERFORMANCE

Jon Clenton is an Associate Professor of Applied Linguistics at Hiroshima University. His research interests relate to vocabulary acquisition in terms of testing, fluency development, English as an Additional Language, and the exploration of learner variation according to first language background.

References

Aizawa, K. (2006). Rethinking frequency markers for English-Japanese dictionaries. In M. Murata, K. Minamiide, Y. Tono, & S. Ishikawa (Eds.), *English lexicography in Japan* (pp. 108-119). Tokyo: Taishukan.

Aizawa, K. & Iso, T. (2007). Estimating word difficulty: the divergence from frequency levels. *Annual Review of English Language Education in Japan (ARELE), 18,* 111-120.

Anderson, R. C. & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews.* (pp. 77-117). Newark, DE: International Reading Association.

Cobb,T. (n.d.) Compleat Lexical Tutor. http://www.lextutor.ca/

Cronbach, L. J. (1942). An analysis of techniques for diagnostic vocabulary testing. *Journal of Educational Research, 36*, 206−217. doi:10.1080/00220671.1942.10881160

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 knowledge.* (pp. 116-133). Cambridge: Cambridge University Press.

Fitzpatrick, T. & Clenton, J. (2010). The Challenge Of Validation: Assessing the Performance of a Test of Productive Vocabulary. *Language Testing, 27*, 537-554. doi:10.1177/0265532209354771

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

Flower, L. & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*, 365-387. doi:10.2307/356600

Fulcher, G. (2013). *Practical language testing.* Abingdon: Routledge.

Henriksen, B. (1999). Three Dimensions of Vocabulary Development. *Studies in Second Language Acquisition, 21*, 303-317. doi:10.1017/S0272263199002089

Horst, M., & Collins, L. (2006). From faible to strong: How does their vocabulary grow?. *Canadian Modern Language Review, 63,* 83-106. doi:10.1353/cml.2006.0046

Jiménez Catalán, R. M. & Moreno Espinosa, S. (2005). Using Lex30 to measure the L2 productive vocabulary of Spanish primary learners of EFL. *VIAL, Vigo International Journal of Applied Linguistics, 2,* 27-44.

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

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

Laufer, B. & Nation, I. S. P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics, 16,* 307-322. doi:10.1093/applin/16.3.307

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

Laufer, B. & Nation, I. S. P. (1999). A Vocabulary-Size Test of Controlled Productive Ability. *Language Testing, 16,* 33-51. doi:10.1191/026553299672614616

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and Strength: Do We Need Both To Measure Vocabulary Knowledge? *Language Testing, 21,* 202-226. doi:10.1191/0265532204lt277oa

Leech, D. (1994). Problematic ESL Content Word Choice in Writing: A Proposed Foundation of Descriptive Categories. *Issues in Applied Linguistics, 5*(1). Retrieved from http://escholarship.org/uc/item/3r55x677

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*, 1–38. doi:10.1017/S0140525X99001776

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams. (Eds.)*, Performance and competence in second language acquisition* (pp. 35–53). Cambridge: Cambridge University Press.

Meara, P. & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short texts. *Prospect 16*(3), 5−19.

Meara, P. & Fitzpatrick, T. (2000). Lex30: An Improved Method of Assessing Productive Vocabulary in an L2. *System, 28*, 19-30. doi:10.1016/S0346-251X(99)00058-5

Meara, P. & Jones, G. (1987). Tests of Vocabulary Size in English as a Foreign Language. *Polyglot*, 8 (1): C1-E14.

Meara, P. & Milton, J. (2003). X_Lex, the Swansea Levels Test. Newbury: Express.

Meara, P. M., & Olmos Alcoy, J. C. (2010). Words as species: An alternative approach to estimating productive vocabulary size. *Reading in a Foreign Language*, *22*(1), 222-236.

Meara, P. & Wolter, B. (2004). V_Links, beyond vocabulary depth. *Angles on the English Speaking World, 4,* 85-96.

Melka, F. (1982). Receptive Versus Productive Vocabulary: A Survey. *Interlanguage Studies Bulletin, 6,* 5-33.

Milton, J. & Hopkins, N. (2005). Aural Lex. Swansea University.

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

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

Nation, I. S.P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*, 59-82. http://dx.doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P. & Beglar, D. (2007). A Vocabulary Size Test. *The Language Teacher, 31*(7), 9-13.

Nation, I. S. P. & Webb, S. A. (2011). *Researching and analyzing vocabulary.* Boston: Heinle, Cengage Learning.

Paribakht, T. S., & Wesche, M. B. (1993). Reading comprehension and second language development in a comprehension-based ESL program. *TESL Canada Journal*, *11*, 9-29.

Paribakht, T. S., & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady & T. Huckin (Eds.) *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 174-200). Cambridge: Cambridge University Press.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*, 513-536. doi:10.1111/1467-9922.00193

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

Richards, J. (1976). The role of vocabulary teaching. *TESOL Quarterly, 10*, 77-89. doi:10.2307/3585941

Walters, J. (2012). Aspects of Validity of a Test of Productive Vocabulary: Lex30. *Language Assessment Quarterly, 9,* 172-185. doi: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,* 33-52. doi:10.1017/S0272263105050023

Webb, S. (2007). The Effects of Repetition on Vocabulary Knowledge. *Applied Linguistics, 28*, 46-65. doi:10.1093/applin/aml048

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

West, M. (1953). *A general service list of English words*. London: Longman, Green & Co.