A Review on Diagnostic Vocabulary Tests

Chunxiu He[1]

[1] Wenzhou Medical University, China

Correspondence: Chunxiu He, Wenzhou Medical University, China

Received: May 4, 2019 Accepted: May 29, 2019 Online Published: May 31, 2019

doi:10.5430/ijhe.v8n3p126 URL: https://doi.org/10.5430/ijhe.v8n3p126

Abstract

Given the practical significance of vocabulary testing in language teaching and the theoretical foundations of developing a vocabulary test, four well-established vocabulary tests are introduced for diagnostic purpose together with their corresponding validation studies, with a focus on the designed purpose, the selection of the items, the presentation of the task and the test response format.

Keywords: vocabulary tests, frequency lists, validation studies

1. Introduction

In EFL context, clear-cut vocabulary learning goals are usually set for students in different phases in national teaching syllabi and quite a large proportion of classroom time is allocated to vocabulary teaching as well, regarding the fundamental position of vocabulary in language learning. Yet many students in general have no clear judgement about their vocabulary abilities due to the lack of knowledge of the testing approach. In vocabulary research field in China, researchers seem to pay no adequate attention to the existing vocabulary tests but turn to use self-designed ones by means of a spaced sampling procedure (Lin, 2003; Wang, 1998) (Note 1). Given that, well-established vocabulary tests should be brought closer to EFL learners and teachers for the pedagogical use and hopefully for related research in China. We finally narrow down to four tests that could be used for diagnostic purpose, namely, the Vocabulary Levels Test (Nation, 1983, 1990; Schmitt et. al., 2001), the Vocabulary Size Test (VST) (Nation & Beglar, 2007), the Productive Levels Test (Laufer & Nation, 1999), and the Lexical Frequency Profile (Laufer & Nation, 1995). Previous research on the validation of the tests are to be included in this paper to check their validity in finding out receptive and productive vocabulary knowledge on the part of a learner, and to explore the pedagogical implication.

2. Practical Significance of Vocabulary Testing

In the design of a language teaching program, it is important that teachers are able to get a grip of the state of learners' vocabulary knowledge. But the fact is, teachers in most cases do not have a justifiable idea of what vocabulary learners already know (Nation, 2011). Teachers lack good intuitions about learners' vocabulary knowledge. According to Biemiller and Boote (2006), about 40% of the words teachers taught are actually known words to learners.

Research also indicates that many students in EFL context may learn vocabulary in an extremely inefficient way, mainly reflected in their inadequate mastery of the high frequency words after several years of study. In a study measuring knowledge of the General Service List (West, 1953), Nurweni and Read (1999) find that after six years of formal English language instruction, first year university students in Indonesia know approximately 60% of the most frequent 1,000 word families, and 37% of the second 1,000 word families. Moreover, the importance of Academic Word List (Coxhead, 2000) as a learning focus for academic study is not known to the learners either. Since learning strategies function effectively to those words learners pay special attention to (Gu, 2010: 115), learners would not apply a range of learning strategies to the words they consider as unimportant. Thus the inactivation in cognitive processing leads to inefficiency in the mastering of "important" words. Therefore, learners need to possess the knowledge about their vocabulary abilities, which is supposed to be facilitated by the guidance from the teacher. However, teachers may believe that the course book will do a satisfactory, if not an excellent, job in planning vocabulary learning since it is written abiding by principled curriculum goals in syllabi. Sadly, the course book sometimes is not a reliable source of vocabulary lists. "Language courses often contain a mixture of useful vocabulary and vocabulary that by no means represents the best choice for those learners." (Nation, 2011)

To sum up, the lack of knowledge about vocabulary ability from both the students and teachers calls for application of diagnostic vocabulary tests in teaching and learning. Only by that can the teacher follow up with other steps such as choosing appropriate instructional path and designing learning tasks for learners.

3. Theoretical Foundations of Vocabulary Testing

Frequency-based studies show very strikingly that some words are much more instrumental than others. Thus, in setting vocabulary learning goals, one must work out how many really useful words learners need to know. By analysing various kinds of tests using 1,000-word family lists made from the British National Corpus (BNC), Nation (2006) deduces that between 3,000 to 4,000 word families (Note 2) are needed to get 95% text coverage, and between 6,000 and 9,000 word families are needed to gain 98% coverage (Table 1). Research shows that 95% coverage may be sufficient for spoken narrative texts (van Zeeland and Schmitt, 2012) and 98% represents a manageable amount of unknown vocabulary, for instance, one word in 50 or one word in every five lines (assuming 10 words per line) (Nation, 2013).

Table 1. English vocabulary sizes (word families) needed to get 95% and 98% coverage (including proper nouns) of various kinds of texts (Nation, 2006)

Texts	95% coverage	98% coverage	Proper nouns
Novels	4,000	9,000	1-2%
Newspapers	4,000	8,000	5-6%
Children’s movies	4,000	6,000	1.5%
Spoken English	3,000	7,000	1.3%

Vocabulary based on frequency levels comes into three types: high-frequency words, mid-frequency words, and low-frequency words. The classic list of high-frequency words is Michael West’s (1953) General Service List, which contains around 2,000 word families including function words and many content words. Schmitt and Schmitt (2014) argue for having a 3,000-word family high-frequency vocabulary list. Mid-frequency words have a coverage of 6,000 to 7,000 (depending on 2,000 or 3,000 word families for high-frequency list), starting from 2,001 or 3,001 to 8,000 or 9,000. They are largely general-purpose vocabulary. Together with high-frequency words, they represent the amount of vocabulary needed to deal with English without the need for outside support. Beyond the frequency of 8,000 or 9,000, words are recognized belonging to low-frequency list. These words consist of technical terms for various subject areas and words that are seldom applied in the use of the language.

Besides frequency, words can also be classified according to the areas that they serve for, normally, general service words, academic words and technical words. Suggested by the name, general service words are for general use. Academic words are words that students often encounter and use in academic texts in the subject areas. The best-known list of academic words is the Academic Word List (Coxhead, 2000), 570 headwords based on a 3,500,000-token corpus of academic English from 4 disciplines of Arts, Science, Law and Commerce, totally covering 28 subject areas. Knowing the 2,000 high-frequency words and the AWL will give close to 90% coverage of the running words in most academic texts (Nation, 2013). For learners studying English for academic purposes, it prioritizes learning that academic vocabulary and the first 3,000 word families are so important groups of vocabulary. Technical words are closely related to the content of a particular discipline. They are mainly for language for specific purposes.

4. Diagnostic Vocabulary Tests

A variety of vocabulary size measures or tests have been developed mainly during the last two or three decades. The tests are generally based on stratified samples of words from frequency lists and assessed in the test score (Wesche & Paribakht, 1996). In introducing vocabulary tests, the designed purpose of the tests needs to be described, and characteristics as well such as (a) the way about the selection of the target vocabulary items; (b) the nature of the task presented to the test-taker; (c) the test response format used; and (d) criteria for judging open-ended responses (if there are) (Wesche & Paribakht, 1996). Read (1993:355-357) notes that vocabulary tests vary along several dimensions in addition to the focus on breadth versus depth. These dimensions are: (a) simple to more complex formats, (b) verifiable responses versus self-report, and (c) contextualization vs. isolation of test items.

The following will present a general picture of four tests for diagnostic purpose in pedagogy: the Vocabulary Levels Test, the Vocabulary Size Test, the Productive Levels Test, and the Lexical Frequency Profile.

4.1 The Vocabulary Levels Test

The Vocabulary Levels Test (VLT) (Nation, 1983, 1990; Schmitt et. al., 2001) is the most applauded vocabulary test currently available for pedagogic use. It is a tool to measure test-takers' written receptive word knowledge in five frequency bands. The first 2,000 word frequency levels take words from West's (1953) GSL, and the 3,000, 5,000, and 10,000 word-frequency bands have those from the lists constructed from Thorndike and Lorge (1944) and Kucera and Francis's (1967) frequency criteria (as cited in McLean & Kramer, 2015). The primary purpose of the test is to diagnose learners' mastery of the most frequent vocabulary to assign appropriate learning materials. Mastery of the level is defined on the basis of correctly 29 or more answers for the 30 items in the corresponding section, since 98% coverage is needed for easy comprehension of written materials (Nation, 2013). As a diagnostic test, it is proved to be very helpful in directing attention to the deserved levels of vocabulary from the teacher and the learner. Items of the VLT are like the following, in which the learners have to match three out of the six words on the left with the meanings listed on the right:

1. original
2. private _____ complete
3. royal _____ first
4. slow _____ not public
5. sorry
6. total

Figure 1. Example of the VLT format (Nation, 1990)

On the website http://www.lextutor.ca/tests, there are already three vocabulary levels tests, namely, Levels Test 2k-10k by Nation (1990), VLT 2k-10k by Schmitt, Schmitt and Clapham (2001) and New VLT by McLean and Kramer (2015)[2]. Take the Vocabulary Levels Test (Nation, 1990) as an example, there are five levels of vocabulary at 2,000, 3,000, 5,000, the University Word Level (UWL) and 10,000, each having 6 sets of items from which 3 out of 6 words have to be matched with their corresponding meanings. Every eighteen items therefore represent the whole of the corresponding frequency level. After the completion of the answering, the website gives a score on the percentage of correct answers out of the tested items and gives a hint on the problem sets if wrong answers take place in the levels. The 2001 version consists of 30 items for each of the five levels in a multiple matching format, in which UWL is changed to AWL on behalf of academic vocabulary.

Different versions have been widely applied in both assessment and research, and quite a few articles have been published to validate the test (Read, 1988; Schmitt et. al. 2001). Read (1988) find the test reliable and that subject scores tend to fall into an implicational scale, for instance, knowing lower-frequency words usually entails the knowledge of higher-frequency ones. Schmitt, Schmitt and Clapham (2001) combine the original four versions into two versions, each with 10 clusters in the five sections. The two versions are tested by a total of 801 subjects in 13 groups in five countries, with the exception of 56 subjects in Group 3, all of whom sit for the test and an interview to explore how closely responses on vocabulary test match the measured lexical knowledge. The individual item works independently well in the test since vocabulary is learned as separate units. The reliability indices (Cronbach's alpha) for all of the Levels sections are high, ranging from .915(for 10k section) to .958 (for academic section). The test through factor analysis is suggested to be unidimensional; personal interviews also indicate that examinees accept the test and that answers on the test could reflect their underlying lexical knowledge. (Schmitt, et al., 2001) It is also found that the items appear to distinguish well between better and weaker learners and thus the test could be used as placement test as well to place students quickly into ability groups based on the vocabulary knowledge.

Since it is a test of receptive vocabulary knowledge, the VLT scores can only indicate the extent to which test-takers know the form and meaning of words rather than the degree to which they can understand or use vocabulary. The role of partial knowledge in vocabulary testing should be considered in interpreting the test score. The knowledge of a word base does not presuppose knowledge of its derivatives or inflections" (Beglar & Hunt, 1999:147) Besides, guessing behaviour may distort results in that there is a 17% chance of correct blind guessing (Kamimoto, 2008; Webb, 2008; as cited in Kremmel & Schmitt). In addition, item interdependence could be a problem in the multiple matching format as, when students answer the items, the decrease of available answer choices reduces the difficulty of finding out the remaining answer (McLean & Kramer 2015). Finally, when applying versions of bilingual tests among learners of different L1, a point worthy of noting in interpreting the test result is that scores taken from the bilingual version of the VLT may display up to 10% higher for lower level learners than on the monolingual versions (Nation & Webb, 2011).

4.2 The Vocabulary Size Test

A more recent test of vocabulary size frequently used is the Vocabulary Size Test (VST) (Nation & Beglar, 2007). The VST is designed to measure written receptive vocabulary knowledge, estimating a total number of the test-takers' vocabulary size. It is a proficiency test but the estimate can also be used to diagnose long-term vocabulary growth and set new vocabulary learning goals. The test is based on word family frequency estimates derived from the spoken subsection of the BNC (Nation, 2006). As Nation and Beglar (2007:10) elaborate, the word family is chosen because learners beyond a minimal proficiency level are assumed to have some control of word building devices and are able to see a formal and a meaning relationship between regularly derivatives of a word family. There are two versions of the test, a 14,000 version containing 140 multiple-choice items and two parallel 20,000 versions containing 100 multiple choice items. Take the former as example, 10 items for each of the fourteen 1,000 word family frequency lists from the Corpus are sampled in order. It is believed that frequency level is directly related to the probability of being known. That is, items in the first 1,000 level are the most likely to be known and those in the 14th 1,000 least likely. Both the 14,000 and 20,000 versions are available at http://www.victoria.ac.nz/lals/about/staff/paul-nation and the online versions for the former also at http://www.lextutor.ca/tests/levels/recognition/ and http://my.vocabularysize.com.

The test presents the target word with a decontextualized sentence by offering four alternatives of the meaning in multiple choices. The total score needs to be multiplied by 100 for the 14,000 version and 200 for the 20,000 version to calculate their total receptive vocabulary size for reading. The result reports the percentage of learners' knowledge of words at each level (partial knowledge is also valued here) and a total number of the estimated vocabulary size and thus the test can be used as a diagnostic test of receptive vocabulary knowledge. Nguyen and Nation (2011) and Karami (2012) suggest that every level of the test should be covered to avoid a considerable underestimation and get a more valid estimate of learners' vocabulary sizes, though some other research advocate that students only need to take the test two levels above their ability (McLean & Kramer, 2015) for an increasing accuracy of the VST results. The following is a sample item from the 5th 1,000 word level.

1. miniature: It is a miniature.

a a very small thing of its kind

b an instrument for looking at very small objects

c a very small living creature

d a small line to join letters in handwriting

Figure 2. Example of the VLT format(Nation & Beglar, 2007)

There is some evidence in support of the validity of the test. In one of the well-designed studies, Beglar (2010) provides validity evidence to the VST by identifying its construct validity in content, substantive, structural, generalizability validity from Messick's validity framework and investigating responsiveness and interpretability. Rasch-based approach to instrument validation is applied to assess the dimensionality of the instrument. It is found that the VST can distinguish learners with widely differing levels of written receptive vocabulary knowledge and provide a sufficient number of frequency levels for measuring learners' lexical acquisition over long periods of time (p.107). Ten items per level are sufficient enough to estimate the test-taker's lexical knowledge with a high degree of precision and the number of items for each level could be substantially reduced, for example, to five. The results from Rasch analysis support that the test displays a high degree of psychometric unidimensionality, i.e, measuring only the written receptive vocabulary knowledge but no other things. The test also performs consistently and reliably with changes in the gender of the subjects, versions of different item numbers and learners of various proficiency levels with a high degree of generalizability. A high degree of responsiveness is indicated in distinguishing persons into levels of ability and its potential to measure changes in lexical knowledge over long periods of time. It suggests that the 14, 000 word frequency level is enough to measure the written receptive lexical knowledge for ESL/EFL learners as knowledge of the most frequent 14,000 words together with proper nouns could account for over 99% of the running words in written and spoken text. The study provides strong evidence for the validity of the test.

A number of bilingual versions of the Vocabulary Size Test have been developed including Arabic, Gujarati, Russian, Korean, Vietnamese, Mandarin and Japanese (http://www.victoria.ac.nz/lals/about/staff/paul-nation). In bilingual versions, the target words are in English and the choices are in learners' L1. The main impetus for the development of such bilingual versions, where there is only equivalence between the tested word and its translation in L1, is the avoidance of the complexity of the choices in explaining the word meaning in L1 and of the extra burden on test takers' grammatical knowledge and reading skill for understanding the choices. The bilingual ones, centred on a single construct of word knowledge underlying the test, have a high level of validity and reliability in distinguishing between different proficiency levels and they are less challenging and more time-efficient than the monolingual version (Nguyen & Nation, 2011; Karami, 2012; Wang & Du, 2014). It would be of great diagnostic use to employ different test forms to measure groups of test-takers' lexical knowledge, especially the progress in learning over time.

4.3 The Productive Levels Test

The Productive Levels Test (PLT) is a more appropriate diagnostic measure of controlled productive vocabulary knowledge, the ability to use a word when compelled to do so. The overall structure of the test is modelled on the Vocabulary Levels Test (Nation, 1983; 1990). Laufer and Nation (1999) modify the VLT format to a test of controlled productive vocabulary ability in the form of sentence completion item type like the following. It prompts test-takers to produce predetermined target words by giving a sentence context or a definition with a clue of the beginning letters of the target words.

The garden was full of fra__ flowers.

Figure 3. Example of the Productive Levels Test (Laufer & Nation, 1999)

There are three parallel test versions of the VLT at http://www.lextutor.ca/tests/levels/productive/, version A, B and C. Take Version A and B for example, each of the test attempts to elicit 18 target words from each of the 2,000, 3,000, 5,000, University Word List (UWL), and 10,000 word levels, using the items from the original Levels Test. For example, if a test-taker gets correct answers for 16 out of 18 items from a certain level, the test will report 88% mastery of the level. As a diagnostic test, the online version also gives suggestions that if the result is below 50%, work on both the first and second 1,000 frequency lists is called on, and if it is between 50-83%, work on the second 1,000 list instead.

In Laufer and Nation (1999), the results of the validation clearly show the gradual mastery of the successive frequency levels of the test as proficiency increases, indicating it is a valid measure of vocabulary growth. The study also offers a mixed version based on the three versions. The equivalence study of the four versions of the test reveals that Version C, a combination of Version A and B that excludes cognates in French, has a satisfactory reliability at four levels of 2,000, 3,000, 5,000 and 10,000 and discriminates between learners of different proficiency levels.

According to Meara and Fitzpatrick (2000), the test is effective mainly for learners with a limited vocabulary size. It can easily identify what the test-takers do not know, but it is not so good at discovering the full extent of what they do know due to the limitation to the choice of the target word. The vocabulary produced by a learner tends to be context-specific, so the true size or range of the learner's productive vocabulary is difficult to be calculated from any small sample. It is also difficult to devise simple tasks to make reasonable estimates for large quantities of vocabulary. Still as a diagnostic test instead of a proficiency test, the PLT looks effectively at learners' productive vocabulary knowledge when testing the knowledge about form-meaning connection.

4.4 The Lexical Frequency Profile

The ability to use a word at one's free will is referred to as free productive ability. This type of vocabulary use is measured by the Lexical Frequency Profile (LFP) (Laufer & Nation, 1995). The test-takers are asked to write 300-or more word essays on a given subject (Laufer & Nation, 1995; Gu, 2010) to diagnose their free control of words of frequency levels. By inputting the essay to a special LFP computer program, calculation is done in seconds to show the percentage of words a learner uses at different vocabulary frequency levels in a composition. It is based on the belief that a test-taker's productive vocabulary size could be deduced from the percentage count of infrequent words he uses in the composition. The frequency levels can be selected from and composed of words at the first 1,000, the second 1,000, the UVW and the 'not-in-the-lists' words. The profiler can also be converted into a more condensed profile of the percentage of the first 2,000 words, the University Word List and 'not-in-the-lists' based on the proficiency level of the subjects. Currently, the special computer program for the LFP, Range, is available at Vocabprofile on Tom Cobb's Web site (http://lextutor.ca/vp/) and Paul Nation's homepage (http://www.victoria.ac.nz/lals/staff/paul-nation). The profile can be calculated for tokens, all words in the composition, for types, different words in the composition, and for word families. Laufer (2012) proposes that the profile reported in tokens could show text difficulty best since repeated occurrences of the same unfamiliar word contribute to comprehension difficulty. However, Laufer and Nation (1995) suggest that the LFP could also measure the quality of lexis in the writing by calculating percentage of types rather than tokens which counts repeated occurrences of the same word.

In Laufer and Nation's study (1995), the subjects are three groups of low-intermediate learners of English, the first-semester Israeli high school graduates and the end-of-two semesters group. All of the students are asked to write two compositions of 300-350 words length on some general topics. Mean percentage and standard deviation at different proficiency levels show that the less proficient students make more use of the first 1,000 most frequent words and they show a similar tendency in the use of the second 1,000, while the more proficient students have a larger percentage in the use of UWL and the 'not-in-the-lists' words. Comparison of any of the two compositions from the same learner in the three groups show that none of the differences between the two essays are significant, showing that the LFP is stable between two compositions by the same person. Since the first 1,000 words are composed of almost all the function words and the most basic lexical words, which are basic units of the written expression, they could not exhibit developed lexicon. The significant differences emerge with the more sophisticated vocabulary like the UVW words and 'not-in-the-lists' words.

The LFP "provides similar stable results for pieces of writing by the same person, and discriminates between learners of different proficiency levels" (Laufer & Nation, 1995: 319). It correlates well with the Productive Levels Test. The main strength includes a test of lexical richness at different levels, laying little emphasis on grammar but lexis. Though the test has two shortcomings in recognizing words -- inability to distinguish homonyms and multiword (Laufer, 2012), the LFP is shown to be a reliable and valid measure of lexical use in writing in learning. As Laufer (2005) claims that the LFP is helpful in telling the proportion of frequent and non-frequent words the learners choose to use in writing rather than offering an estimate of productive vocabulary size. It offers evidence that lexical knowledge and lexical use develop through different pathways.

Meara and Fitzpatrick (2000) argue that the constraint of the topic selection in the task reflects the context-limited property of the test to some degree, but it is much weaker than the PLT in this aspect. It may be not a cost-effective way of eliciting vocabulary, in that at least two 300-word essays from the test-takers are needed in order to obtain stable vocabulary size estimates, which would take about two hours of class time, making it difficult to implement for research, but the presentation of test results is amazingly efficient in categorizing words into corresponding frequency levels to provide a detailed picture of vocabulary use of a learner. Therefore, it is highly recommendable to have a diagnostic examination of learners' vocabulary use in writing with the LFP.

5. Conclusion

The four types of vocabulary measures serve for diagnostic purpose of testing the receptive or productive vocabulary knowledge in a learner. It is assumed that learners' vocabulary size as measured by a vocabulary test is to be expected in learners' productive use of the language (Laufer & Nation, 1995: 319), but the relation between them needs to be further explored. The test purpose and format are summarized in the Table 2 and all of the tests are accessible at Tom Cobb's http://www.lextutor.ca.

Table 2. Summary of the four vocabulary testing measures

Tests	Testing Purpose	Testing Format
the VST	receptive vocab. size	multiple choice
the VLT	receptive vocab. size	matching
the PLT	controlled productive vocab. ability	word completion
the LFP	free productive vocab. use	essay writing

The research field still has quite a few long-standing vocabulary testing measures such as the Eurocentres Vocabulary Size Test (Meara & Jones, 1988) and the Word-Associates Test (Read, 1993), and a lot more are coming up to measure different aspects of vocabulary knowledge and ability, for instance, the New Vocabulary Levels Test (NVLT) (Note 3) expanding the VLT, the Academic Word Levels Test (AWLT) assessing learners' knowledge of the different sublists of AWL, the Word Part Levels Test (WPLT) and the Guessing from Context Test (GCT) testing learners' vocabulary learning ability (Webb & Sasao, 2013). Research still calls for the development of the testing measures (Note 4), since none could be regarded as perfect. Besides, these tests, old and new, could be combined in an integrative manner to report a comprehensive picture of learners' vocabulary knowledge in different aspects for instructional purpose (for an example, see Ishii, 2009). For research purpose, the new testing measures can be incorporated with the earlier tests to crosscheck or validate each other. All in all, it would certainly shed lights on vocabulary teaching and learning if these measures are appropriately made use of in vocabulary knowledge/ability testing and use. And thus it is of urgent necessity that these vocabulary measures should be promoted among English teachers and learners first and foremost, and then chances are there to apply and develop them in practice.

Acknowledgements

The research is sponsored by Wenzhou Medical University Project *Blended Vocabulary Teaching Based on Blackboard* (No. YBKG201718) and Zhejiang Education and Science Planning Project *Action Research on Vocabulary Testing and Strategy Learning: the developing of Vocabulary Course on Blackboard* (No. 2016SCG179).

References

Beglar, D. & Hunt, A. (1999). Revising and validating the 2000 Word Level and University Word Level Vocabulary Test. *Language Testing, 16*(2), 131-162. https://doi.org/10.1177/026553229901600202

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

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *27*(1), 101–118. https://doi.org/10.1177/0265532209340194

Biemiller, A. & Boote, C. (2006). An effective method for building meaning vocabulary in primary grades. *Journal of Educational Psychology, 98*(1), 44-62. https://doi.org/10.1037/0022-0663.98.1.44

Coxhead, A. (2000). A new academic word list. *TESOL quarterly*, *34*(2), 213-238. https://doi.org/10.2307/3587951

Feng, Y-F. (2003). A comparative study of vocabulary learning strategies in English majors from different grades. *Foreign Language World, 2*, 66-72(in Chinese with English abstract)

Gu, Y. (2010). Learning strategies for vocabulary development. *Reflections on English Language Teaching, 9*(2), 105-118.

Ishii, T. & Schmitt, N. (2009). Developing an integrated diagnostic test of vocabulary size and depth. *RELC Journal*, *40*(1), 5-22. https://doi.org/10.1177/0033688208101452

Karami, H. (2012). The Development and Validation of a Bilingual Version of the Vocabulary Size Test. *RELC Journal*, *43*(1), 53–67. https://doi.org/10.1177/0033688212439359

Kremmel, B. & Schmitt, N. Vocabulary Levels Test. retrieved on June 3, 2016 at http://scholar.google.co.nz/scholar?q=Kremmel%2C+B.%2C+%26+Schmitt%2C+N.+Vocabulary+Levels+Test.&btnG=&hl=en&as_sdt=0%2C5

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

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

Laufer, B. (2005). Lexical Frequency Profile: From Monte Carlo to the Real World A Response to Meara. *Applied Linguistics, 26*(4), 582-588. https://doi.org/10.1093/applin/ami029

Laufer, B. (2012). Lexical Frequency Profiles. The Encyclopedia of Applied Linguistics, Edited by Carol A. Chapelle. Published 2013 by Blackwell Publishing Ltd. https://doi.org/10.1002/9781405198431.wbeal0692

Lin, M. (2003). Chinese non-English majors' strategies in vocabulary learning. *Teaching English in China*, *26*(2), 56-60.

McLean, S. & Kramer, B. (2015). The creation of a new vocabulary levels test. *SHIKEN*, *19*(2), 1-11. https://doi.org/10.1177/1362168814567889

Meara, P. & Jones G. (1988). Vocabulary Size as Placement Indicator. http://eric.ed.gov/?id=ED350829

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

Nation. I. S. P. (1983). Testing and teaching vocabulary. *Guidelines*, *5*(1), 12-25.

Nation, I. S. P. (1990). *Teaching and Learning Vocabulary*. Rowley, MA: Newbury House.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening?. *Canadian Modern Language Review, 63*(1), 59-82. https://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. (2011). Research into practice: Vocabulary. *Language Teaching, 44*(4), 529-539. https://doi.org/10.1017/S0261444811000267

Nation, I. S. P. & Webb, S. (2011). *Researching and Analysing Vocabulary*. Boston: MA. Heinle.

Nation, I. S. P. (2013). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139858656

Nguyen, L. T. C. & Nation, I. S. P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. *RELC Journal, 42*(1), 86-99. https://doi.org/10.1177/0033688210390264

Nurweni, A. & Read J. (1999). The English Vocabulary Knowledge of Indonesian University Students. *English for Specific Purposes, 18*(2), 161–175. https://doi.org/10.1016/S0889-4906(98)00005-2

Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC journal, 19*(2), 12-25. https://doi.org/10.1177/003368828801900202

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10*(3), 355-371. https://doi.org/10.1177/026553229301000308

Read, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511732942

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

Schmitt, N. & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, 47*(4), 484-503. https://doi.org/10.1017/S0261444812000018

van Zeeland, H. & Schmitt, N. (2012). Lexical coverage and L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*. https://doi.org/10.1093/applin/ams074

Wang, W-Y. (1998). Beliefs, strategies and English vocabulary retention. *Foreign Language Teaching and Research, 1*, 47-52(in Chinese with English abstract).

Wang, Y. & Du, W. (2014). Study on the Validity of bilingual Mandarin Version of Vocabulary Size Test. *International Journal of English Linguistics, 4*(6), 113-117. https://doi.org/10.5539/ijel.v4n6p113

Webb, S. A. & Sasao, Y. (2013). New directions in vocabulary testing, *RELC Journal, 44*(3), 263-277. https://doi.org/10.1177/0033688213500582

Wesche, M. & Paribakht, T.S. (1996). Assessing second language vocabulary knowledge: depth versus breadth. *Canadian Modern Language Review, 53*(1), 13-40. https://doi.org/10.3138/cmlr.53.1.13

West, M. (1953). *A General Service List of English Words*. London: Longman, Green & Co.

Wu, X. & Wang, Q. (1998). Vocabulary learning strategies by non-English majors. *Foreign Language Teaching and Research, 1*, 53-57(in Chinese with English abstract)

Xue, G. & Nation, I. S. P. (1984). A university word list. *Language learning and communication, 3*(2), 215-229.

Notes

Note 1. In the study of vocabulary learning strategies, Wu & Wang (1998) used Level 3,000 in VLT (Nation, 1900) and Feng (2003) employed five levels from VLT, but neither elaborated the test in their studies.

Note 2. A word family is defined as a headword plus the inflected and closely related derived forms.

Note 3. Despite the connection in the name, New VLT actually should be recognized as an independent test from VLT, since it assesses the first five 1, 000-word frequency levels of the BNC and the AWL and it also uses a different format of multiple choice. (see McLeon and Kramer, 2015)

Note 4. When the article is written up, CATTS (Computer Adaptive Test of Size & Strength) (Laufer & Levitzky-Aviad, 2016)--a frequency-based measure testing both receptive and productive vocabulary knowledge has been included on the website.