Vocabulary Learning and Instruction
Volume 1, Issue 1, August 2012
http://vli-journal.org

Four Empirical Vocabulary Test Studies in the Three Dimensional Framework

Masamichi Mochizuki
Reitaku University
doi: http://dx.doi.org/10.7820/vli.v01.1.mochizuki

Abstract

In this paper I would like to briefly overview vocabulary testing literature and discuss the four empirical studies conducted by Jeffrey Stewart, Rie Koizumi, Aaron Batty, and Tatsuo Iso, after placing them in the framework of the three dimensions of vocabulary knowledge: size, depth, and lexical accessibility.

Keywords: three dimensions of vocabulary knowledge: size, depth, and lexical accessibility; passive/receptive vocabulary knowledge; active/productive vocabulary knowledge.

1 Tests of Vocabulary Size

The first dimension of vocabulary knowledge (Daller, Milton, & Treffers-Daller, 2007) is size or breadth. Researchers have developed a variety of measures to assess vocabulary size. Initially a number of attempts were made to estimate English native speakers' vocabulary sizes by sampling words from a dictionary and challenging test-takers to give definitions of those words. Subsequently more sophisticated methods of testing vocabulary have been proposed: The Vocabulary Levels Test (VLT) (Nation, 1990, 2001; Schmitt, Schmitt, & Clapham, 2001); The Eurocentre Vocabulary Test (Meara & Jones, 1990); The Mochizuki test (Mochizuki, 1998); The Vocabulary Size Test (Nation & Beglar, 2007). All these tests are considered to measure test-takers' receptive vocabulary sizes.

On the other hand, very few attempts have been made to assess test-takers' productive vocabulary sizes. Laufer and Nation (1999) developed a productive version of the VLT which challenged test-takers to complete a word in a sentence context whose first few letters were given as a cue. It is possible to estimate test-takers' productive vocabulary sizes with the Productive Levels Test, even though the VLTs were intended to diagnose the frequency levels learners needed to study (Nation, 1990). Another attempt to measure productive vocabulary size is Lex 30 (Meara & Fitzpatrick, 2000). In this test, participants give as many associations as they can think of from 30 stimulus words. The number of associations beyond the most frequently used 1000-word level is counted and used as an indicator of the test-taker's productive vocabulary size. Laufer and Goldstein (2004) proposed to measure vocabulary knowledge *strength* because learners differ in their vocabulary knowledge along the two axes, active/passive and recall/recognition. Active recall knowledge corresponds to productive vocabulary in that a learner is able to produce a second language word on his or her first language word cue. Passive recognition knowledge may correspond to what multiple-choice vocabulary size tests measure, i.e., receptive vocabulary. Mochizuki (2007) developed a computer program vocabulary test, the J8VST, which measures Japanese English as a foriegn language (EFL) learners'

Please cite this article as: Mochizuki, M. (2012). Four empirical vocabulary test studies in the three dimensional framework. *Vocabulary Learning and Instruction, 1*(1), 44–52. doi: 10.7820/vli.v01.1.mochizuki

vocabulary sizes in four modes: L2 recall; L1 recall; L2 recognition; and, L1 recognition. It is possible to estimate productive vocabulary size by using the J8VST.

There have been several studies that compared L2 learners' receptive and productive vocabulary sizes using the receptive and productive Levels Test (Fan, 2000; Laufer, 1998; Waring, 1997). Webb (2008) criticized these studies for comparing the two dimensions of vocabulary knowledge by using measures that favor the estimation of receptive knowledge. He used recall tests for both measuring receptive and productive vocabulary sizes and found that the scores on the two vocabulary tests were almost the same when a sensitive scoring method was employed, although the scores on the productive vocabulary test were lower than those of the receptive one when strict marking was used.

What is at issue now is that there are no established productive vocabulary size tests that can estimate test-takers' productive vocabulary in a valid, reliable, and efficient manner. An L1 to L2 recall test seems to be a valid and reliable way of testing productive vocabulary but responses must be marked manually. A recall test in a computer program, like the J8VST, can be administered more efficiently than the one marked manually, but it requires an enormous amount of time to develop.

With productive vocabulary measurement in this state, Jeffrey Stewart's study is an epoch-making breakthrough. He has developed a productive vocabulary test, *The KSU Active Multiple-Choice Test*, or "Active MC" test, which requires test-takers to recall an L2 word based on an L1 definition prompt. The second and third letters of the target word as well as a blank for each of the remaining letters are also provided as a hint and to avoid other possibly correct answers. The test-takers simply mark the first letter of the target word on an answer sheet. Stewart found an extremely high correlation (0.93) between this new test and a conventional recall test of the same items. The most appealing feature of this new test is that it can be marked by an optical mark reader. It releases researchers from troublesome manual marking, saving great amounts of time and effort. It will definitely increase the number of studies that address the productive vocabulary of English as a second language (ESL) and EFL learners.

Stewart's Active MC format tests learners' productive vocabulary in the discreet, selective, and context-independent dimensions of the vocabulary test categorization framework proposed by Read (2000). Researchers have also been investigating learners' productive vocabulary in the other poles of the Read's framework: the embedded, comprehensive, and context-dependent dimensions. They have developed indices that are intended to measure the lexical richness, lexical diversity, or lexical density of learners' writing or speech (e.g. Laufer & Nation, 1995). Probably the type-token ratio (TTR) is the best-known measure of lexical diversity though it has the drawback of being affected by text length. Consequently, other indices have been proposed to overcome this defect: the Guiraud index; D; and, the measure of textual lexical diversity (MTLD). Rie Koizumi investigated the reliability of these four measures of lexical diversity when measuring texts of differing lengths. She found that among the four measures, MTLD was least affected by text length. Koizumi also concluded that texts should be at least 100 words long for lexical diversity investigation. These findings hold promise for researchers investigating L2 learners' speaking or writing skill development. They are now able to measure the lexical diversity of learners' speaking or writing performances

at relatively early stages of language development because they require samples of learner production as short as 100 words.

However, it may be argued that researchers should take samples produced on a variety of topics when examining the lexical diversity of learner production, because lexical diversity is likely to vary according to the topic under consideration. For instance, learners can produce a number of different subordinate words on a topic they are familiar with. A learner who loves dogs, for instance, may produce various names like Scotch terrier, Pomeranian, Siberian Husky, etc. on the topic of dogs. If the same learner has little interest in birds, however, she/he will likely keep using the common word *bird* on that topic because of her/his ignorance of bird types. This learner's lexical diversity index of a text on dogs will be much higher than that of a text on birds. Thus, it is important to collect samples of texts produced on a variety of topics in order to show the general development of lexical diversity among learners.

It should also be noted that lexical diversity is but one of the properties of text production: there are other properties that characterize text quality such as cohesion and coherence, organization, grammar, etc. In the case of beginners who produce texts as short as 100 words, it may be argued that lexical diversity does not have a high priority: pronunciation, delivery, and size may be more important in speech; grammar, cohesion and coherence, and organization may be more crucial in writing.

Despite the above, Koizumi's findings still lend strong support for the measures of lexical diversity as a tool for investigating productive vocabulary in the embedded, comprehensive, and context-dependent dimensions.

2 Tests of Vocabulary Depth

We have been talking about the vocabulary size dimension, especially productive vocabulary. Now let us turn to the depth dimension. In my opinion, it was Harold Palmer, often considered the "father of British applied linguistics" (Stern, 1983, p.100), who first drew attention to one aspect of vocabulary depth, collocation. He not only selected 3000 headwords, which he argued together with their commonest derivatives would cover 95% of the contents of all ordinary English texts (Palmer, 1931), but also made a tentative list of English collocations for technicians so that they could apply it to textbook compilation (Palmer, 1933). This shows his understanding of the importance of vocabulary depth knowledge and innovative insight into vocabulary teaching. Other researchers also advanced our understanding of the components of word knowledge in the first half of the 20th century. Along with Palmer, Michael West and Laurence Faucett, both experienced teachers and researchers, and E.L. Thorndike, a statistical linguist, agreed on seven criteria for vocabulary selection for ESL learners: (1) word frequency; (2) structural value (all structure words included); (3) universality (words likely to cause offence locally—excluded); (4) subject range (no specialist items); (5) definition words (for dictionary-making, etc.); (6) word-building capability; and (7) style ("colloquial" or slang words—excluded) (Howatt, 1984, p. 256). Using these criteria they produced the so-called "Carnegie Report" on vocabulary selection in 1936, which West later published as A General Service List of English Words (GSL; West, 1953).

These works by Palmer (1931, 1933) and West (1953) demonstrate that the researchers understood the necessity of teaching vocabulary depth knowledge as well as increasing learners' vocabulary sizes. In the same year as the GSL was published, Dolch and Leeds (1953) emphasized the importance of measuring less common usages of words. They argued that vocabulary tests should measure different senses of a word because understanding the sense a speaker or writer intends to convey is crucial in successful communication. This argument seems, to the best of my knowledge, to be the first to address the necessity to measure vocabulary depth and degrees of learners' developing vocabulary knowledge.

As a method of measuring the degree of a learner's vocabulary knowledge, Dale (1965) proposed a four-stage self-report scale:

Stage 1: "I never saw it before."
Stage 2: "I've heard of it, but I don't know what it means."
Stage 3: "I recognize it in context — it has something to do with . . ."
Stage 4: "I know it."

In this scale learners self-report their knowledge of individual words in four stages. Results reveal how firmly a learner thinks he knows each word. The scale was intended for L1 English learners. Later, Paribakht and Wesche (1993) proposed a similar scale, the Vocabulary Knowledge Scale, for L2 learners with the final stage asking if a learner was able to use the word productively.

There have been a number of attempts to measure some aspects of vocabulary depth knowledge. For example, Ordonez, Carlo, Snow, and McLaughlin (2002) employ definition tasks that they claim elicit not only paradigmatic but also syntagmatic knowledge of a word and thus measure depth of vocabulary knowledge. In investigation of other components of word knowledge, researchers have invented their own measurements, for example: affix and association (Schmitt & Meara, 1997); synonym and collocation (Mochizuki, 2002); antonym, derivation, and collocation (Koizumi, 2005). These are mostly one-time studies that are not replicated or modified in other studies and thus the testing methods are not developed or improved, let alone, standardized.

Among a number of attempts to measure depth of vocabulary the Word Associates Test (WAT) (Read, 1993; 1998) has been most widely employed by researchers (e.g. Qian, 1999, 2002). Although the WAT appears to test synonyms and collocates, one study showed that the test might measure only one dimension; i.e. vocabulary (Schoonen & Verhallen, 2008). This is what Aaron Batty addressed in his study. He showed that the WAT best fitted a bifactor model that presumes the test measures synonym and collocate factors, as well as a general vocabulary factor. This model is based on the assumption that all test items are affected by the general vocabulary factor in addition to half of the items being affected by the synonym factor and the other half by the collocate factor. So Batty revealed that the WAT measures not only the general vocabulary knowledge but also the synonym and collocational aspects of vocabulary.

One thing I am particularly concerned about with the WAT is a high correlation coefficient with a vocabulary size test. Qian (2002) reported a high correlation, 0.88, between his WAT and the VLT (Nation, 1990) scores. This means that the two tests share more than 77% of their variance. This raises doubts on

the validity of the WAT as a measure of vocabulary depth: it may measure the size dimension more than the depth dimension. It might be regarded as another measure of vocabulary size rather than a measure of vocabulary depth. Batty reported that the highest factor loadings were for vocabulary g-factor, with half of the synonym items and only two collocate items loading higher on the corresponding factors than on the g-factor. Because the vocabulary g-factor and the synonym factor are considered to be major components of vocabulary size tests, the WAT can be primarily seen as a vocabulary size test with a minor component of vocabulary depth.

So what makes the WAT a vocabulary size test? In my opinion it is the synonym section. Although it tests different senses of adjectives, it may be argued that learners' knowledge of different senses of polysemous words is highly correlated with their vocabulary size in that both are based on their knowledge of word meaning. Thus, it seems on the surface that the WAT measures the depth aspect of polysemous meanings but it actually measures the vocabulary size dimension.

From this it may be concluded that although Batty's study indicates that the WAT is best fitted with a bifactor model, it would seem more appropriate to regard it as another test of vocabulary size. Therefore, we should make further efforts to create a measure of vocabulary depth. Or should we review the construct of *vocabulary knowledge* as Vermeer (2001) claimed that there was no conceptual difference between breadth and depth measures of vocabulary?

3 Tests of lexical access

The third dimension of vocabulary knowledge is lexical accessibility or fluency. Although lexical accessibility has been vigorously investigated in psycholinguistics, only a limited number of researchers have addressed the issue in applied linguistics (Aizawa & Iso, 2010; Coulson, 2005; Kadota, 2010; Meara, 2005). These studies share a common feature in that they used self-designed computer programs to measure lexical access time instead of established software such as SuperLab. Coulson (2005) used a computer program called Q_Lex to measure lexical access time. This test challenges test-takers to find an English word in a string of letters and measures the time between when a string appears on the computer display and when a test-taker clicks the button to indicate he/she has found the word in the string. Aizawa and Iso (2010) reported their attempts to create a computer program, the Lexical Access Time Test (LEXATT), and found that measuring the time duration while a test-taker is pressing a key is a reliable way of lexical access time measurement. Tatsuo Iso subsequently revised LEXATT as LEXATT2 for more accurate measurement. Kadota (2010) applied the psycholinguistic method of measuring priming effects to measuring lexical access time. He developed a computer program, the Computer-Based English Lexical Processing Test (CELP), which shows a test-taker a prime and then a target word and measures how fast the test-taker judges whether the prime and the target word are semantically related or not. Although studies show that there are significant correlations between CELP scores and English skill test scores (Hase & Shiki, 2011; Nakanishi & Sugiura, 2011), no studies have been conducted to validate CELP as a test of measuring lexical access speed.

In such circumstances LEXATT2 will likely become a standard tool for lexical access measurement for two reasons. First, it is a valid test of lexical

access time. Less proficient test-takers responded more slowly to stimulus words than their more proficient peers and took more time in responding to longer words while proficient test-takers reacted in similar times irrespective of word length. These findings can be seen as evidence of the validity of LEXATT2 for measuring lexical access time. Second, LEXATT2 is extremely practical. Learners can take the test on-line and it requires only 10–15 minutes to complete. So it is a highly practical tool for lexical access time measurement compared with conventional recognition speed tests.

Although LEXATT2 has been validated by Iso, further validation by comparing its reaction times with those of SuperLab would establish its role as a test of lexical access time.

4 Conclusion

The four studies I have examined in this paper had their roles in the framework of vocabulary testing. They have contributed a great deal to the development of each of the vocabulary research fields. I would like to close this paper with suggestions to the four researchers for further studies. First, Jeffrey Stewart could make the most of his productive vocabulary size test by using it to investigate relationships between ESL learners' productive skills and productive (active) vocabulary sizes. Though there are a number of studies investigating relationships between learners' skills and *receptive* vocabulary sizes, very few studies have addressed how their *productive* vocabulary sizes are related to their productive skills. Stewart's new productive vocabulary size test should promote explorations into this area. Second, Rie Koizumi could explore how ESL learners' speeches and writings may be accounted for by different factors such as lexical density, vocabulary size, grammar, cohesion and coherence, and discourse organization. It would be a great discovery to find the extent to which different factors account for productive skills. Third, Aaron Batty could continue examining what the WAT actually measures. Although he found that the WAT fits the bifactor model best, the test may be just another vocabulary size test. He could identify what causes it to behave more like a size test (what would happen if low-frequency words were removed?) rather than a depth test, and then go on to create a valid vocabulary depth test. Finally, Tatsuo Iso could probe relationships between English skills and lexical access time using LEXATT2. It would contribute a great deal to vocabulary acquisition research if it was revealed what role lexical access plays in L2 skill performances.

It is important to further investigate vocabulary testing in the vocabulary knowledge framework so that learners' vocabulary knowledge can be better diagnosed. It is also necessary to try to integrate tests of different dimensions so that results can show test-takers' vocabulary knowledge as a whole and better predict their language proficiency.

References

Aizawa, K., & Iso, T. (2010). *Development of a vocabulary test battery estimating English skills and proficiency: Integrating vocabulary size, organization, and access speed*. Retrived from Mochizuki Masamichi, Reitaku University, Japan. (Research No. 19320084).

Coulson, D. (2005). *Recognition speed for basic L2 vocabulary.* A paper read at the Second JACET English Vocabulary Research Group Conference, Chuo University.

Dale, E. (1965). Vocabulary measurement: Techniques and major findings. *Elementary English*, *42*, 395–401.

Daller, H., Milton, J., & Treffers-Daller, J. (2007). *Modelling and assessing vocabulary knowledge*. Cambridge: Cambridge University Press.

Dolch, E.W., & Leeds, D. (1953). Vocabulary tests and depth of meaning. *Journal of Educational Research*, *47*, 181–189.

Fan, M. (2000). How big is the gap and how to narrow it? An investigation into the active and passive vocabulary knowledge of L2 learners. *RELC Journal*, *31*, 105–119. doi:10.1177/003368820003100205

Hase, N., & Shiki, O. (2011). What the data from CELP test tell us: Relationship between CELP test and reading measurements for fluency. In N. Hase (Chair), *Exploring the roles of lexical processing in second language proficiency: Using the newly developed vocabulary processing test*. Symposium conducted at the JACET 50th Commemorative International Convention, Seinan Gukuin University, Fukuoka, Japan.

Howatt, A.P.R. (1984). *A history of English language teaching*. Oxford: Oxford University Press.

Kadota, S. (2010). *The Interface between lexical and sentence processing in L2: An empirical study of Japanese EFL learners.* Retrived from Kadota Shuhei, Kwansei Gakuin University, Japan. (Research No. 19520532).

Koizumi, R. (2005). 日本人中高生における発表語彙知識の広さと深さの関係 (Relationship between breadth and depth of productive vocabulary knowledge of Japanese junior and senior high school students). *STEP Bulletin*, *17*, 63–80.

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

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

Meara, P. (2005). Designing vocabulary tests for English, Spanish, and other languages. In C. Butler, M.A. Gomez Gonzalez & S. Doval Suarez (Eds.), *The dynamics of language use: Functional and contrastive perspectives* (pp. 271–285). Amsterdam: John Benjamins.

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. (1990). *The Eurocentres vocabulary size test. 10KA*. Zurich: Eurocentres.

Mochizuki, M. (1998). Nihonjin eigo gakushusha no tameno goi saizu tesuto (A vocabulary size test for Japanese learners of English). *IRLT Bulletin*, *12*, 27–53.

Mochizuki, M. (2002). Exploration of two aspects of vocabulary knowledge: Paradigmatic and collocational. *Annual Review of English Language Education, 13*, 121–129.

Mochizuki, M. (2007). *Construction of a vocabulary list for Japanese learners of English and development of a system from analysing educational materials based on large-scale corpora*. Retrived from Aizawa Kazumi, Tokyo Denki University, Japan. (Research No. 6320076).

Nakanishi, H., & Sugiura, K. (2011). Relationships among lexical access speed, WM capacity and speaking skills. In N. Hase (Chair), *Exploring the roles of lexical processing in second language proficiency: Using the newly developed vocabulary processing test*. Symposium conducted at the JACET 50th Commemorative International Convention, Seinan Gukuin University, Japan.

Nation, I.S.P. (1990). *Teaching and learning vocabulary.* New York, NY: Newbury House.

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

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

Ordonez, C.L., Carlo, M., Snow, C., & McLaughlin, B. (2002). Depth and breadth of vocabulary in two languages: Which vocabulary skills transfer? *Journal of Educational Psychology, 94*, 719–728. doi:10.1037/0022-0663.94.4.719

Palmer, H.E. (1931). *Second interim report on vocabulary selection*. Submitted to the Eighth Annual Conference of English Teachers Under the Auspices of the Institutes for Research in English Teaching, Tokyo.

Palmer, H.E. (1933). *Second interim report on English collocations.* Tokyo: Kaitakusha.

Paribakht, T.S., & Wesche, M.B. (1993). Reading comprehension and second language development in a comprehension-based ESL program. *TESL Canada Journal*, *11*, 9–27.

Qian, D.D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review*, *56*, 282–307. doi:10.3138/cmlr.56.2.282

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. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*, 355–371. doi:10.1177/026553229301000308

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunan (Ed.), *Validation in language assessment* (pp. 41–60). Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition*, *19*, 17–36.

Schoonen, R., & Verhallen, M. (2008). The assessment of deep word knowledge in young first and second language learners. *Language Testing*, *25*, 211–236. doi:10.1177/0265532207086782

Stern, H. (1983). *Fundamental concepts of language teaching*. Oxford: Oxford University Press.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics*, *22*, 217–234. doi:10.1017/S0142716401002041

Waring, R. (1997). A comparison of the receptive and productive vocabulary sizes of some second language learners. *Immaculata*, *1*, 53–68.

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, *30*, 79–95. doi:10.1017/S0272263108080042

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