Vocabulary Levels Test

BENJAMIN KREMMEL AND NORBERT SCHMITT

Framing the Issue

The Vocabulary Levels Test (VLT) has been called the nearest thing to a standardized vocabulary test currently available (Meara, 1994, 1996). It was originally developed by Paul Nation in the 1980s (published in Nation, 1990), and subsequently revised by Schmitt, Schmitt, and Clapham in 2001. It is a tool to measure the written receptive vocabulary knowledge, that is mainly the word knowledge required for reading. The VLT assesses this knowledge of learners at four frequency levels of English word families: 2,000, 3,000, 5,000 and 10,000, hence the name "Levels Test." Versions of this test are available freely on the personal websites of Paul Nation (printable, http://www.victoria.ac.nz/lals/about/staff/paul-nation), Tom Cobb (printable and online, http://www.lextutor.ca/tests/), and Norbert Schmitt (printable, www.norbertschmitt.co.uk), the latter two providing the latest revised versions of the test (Schmitt, Schmitt, & Clapham, 2001). Versions of the test can also be found in Nation (1990), Schmitt (2000), and Schmitt (2010).

Each section of the revised VLT consists of 30 items in a multiple matching format. Three items therefore represent 100 words of any particular frequency band. Items are clustered together in 10 groups for this, so that learners are presented in each cluster with six words in a column on the left and the corresponding meaning senses of three of these in another column on the right. Learners are asked to match each meaning sense in the right-hand column with one single word from the left-hand column. Thus, the test asks learners to recognize the form rather than the meaning, that is, the options are words instead of definitions (Schmitt, 2010). As such, the VLT taps the very basic and initial stages of form-meaning link learning. Example items from three of the levels can be seen in Figure 1.

Each cluster targets three words, although some researchers have argued that knowledge of the meaning of the three distractor words is also tested as the test takers need to be familiar with them when they discard them (Read, 1988). Within each level, there is a fixed ratio of word classes to represent the distribution of English word classes. This ratio was 5 (noun) : 3 (verb) : 1 (adjective) in the initial

The TESOL Encyclopedia of English Language Teaching.
Edited by John I. Liontas (Project Editor: Margo DelliCarpini; Volume Editor: Neil J Anderson).

DOI: 10.1002/9781118784235.eelt0499

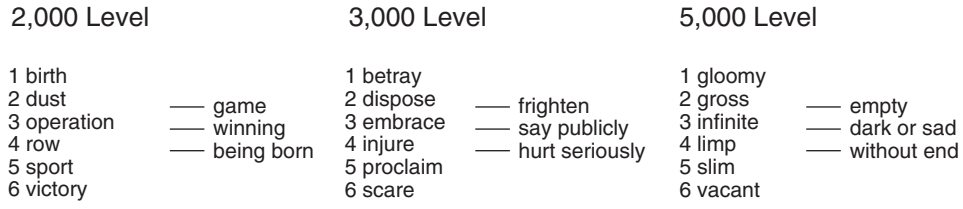

Figure 1 Examples from the Vocabulary Levels Test (VLT) (Schmitt et al., 2001).

version of the VLT (Beglar & Hunt, 1999) and is now 3 (noun) : 2 (verb) : 1 (adjective) in the latest revised versions (Schmitt et al., 2001). Word classes are not mixed within any one cluster. For the 2001 version of the VLT, two parallel test versions are available that have been established to be relatively equivalent (Schmitt et al., 2001, Xing & Fulcher, 2007).

In addition to the four frequency-based levels, the VLT includes test items from the University Word List (UWL) (Xue & Nation, 1984) in the 1990 version and, more recently, the Academic Word List (AWL) (Coxhead, 2000) in the 2001 version. The item sampling of the AWL section, however, is not based on frequency levels, and so it should not be interpreted together with those levels. Nevertheless, the section might be useful as a separate measure for teachers in academic contexts.

Making the Case

The VLT has been established to work well as an instrument for diagnosis, that is, identifying lexical weaknesses at a particular frequency level, and placement, that is, placing students quickly into ability groups based on their vocabulary knowledge (Schmitt et al., 2001; Huhta, Alderson, Nieminen, & Ullakonoja, 2011). Although there are surprisingly few studies published that investigated the validity of the instrument, the VLT seems a very useful tool for teachers for these purposes.

The most comprehensive validation of the revised VLT was undertaken by Schmitt et al. (2001). Their thorough study with 801 EFL learners from different countries, found that reliability was high in their test versions with an increased 30 items per level compared to previous test versions, and that the items appeared to distinguish well between better and weaker learners (Schmitt et al., 2001).

One of the few other validation studies conducted by Read (1988) revealed that an implicational scale can be assumed for the frequency levels. One would generally expect candidates who knew lower-frequency words to also know high-frequency words. That is, the 5,000 level is expected to be more difficult or less well known than the 2,000 or 3,000 level. In the VLT, Schmitt et al. (2001) found that there is indeed a consistent stairstep pattern of mastery of the ascending frequency

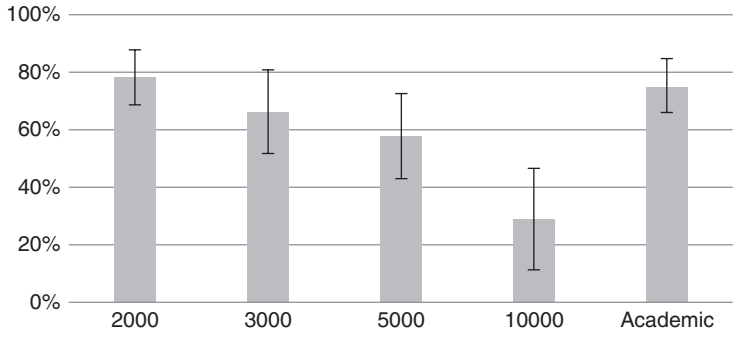

Figure 2 A consistent stairstep pattern of mastery of the ascending frequency levels across a large group of learners.

levels across a large group of learners, which further confirms the functioning of the test (Figure 2).

A comparison of test scores with scores of candidates on a more comprehensive interview measure suggested that the VLT, despite being a selected response format, showed relatively few problems with guessing behavior distorting results (Schmitt et al., 2001). However, Kamimoto (2008) and Webb (2008) suggest there is a 17% chance of learners blind guessing correct responses. Stewart and White (2011) claim that item interdependence, as a function of the multiple matching format, further complicates this issue in the VLT. Since the distractors are words from the same frequency level as the targets, the overestimation in scores due to guessing is variable depending on the proportion of distractors known to a candidate. Their multiple guessing simulations on the VLT found that candidates' scores are generally and consistently inflated by 16–17 points on a 99-item VLT test "until over 60% of words are known, at which point the score increase due to guessing gradually begins to diminish" (Stewart & White, 2011, p. 378).

Beglar and Hunt (1999) cautioned about the interpretation of VLT scores, the sampling of which is based on word family frequency lists. They state that "knowledge of a word's base form does not guarantee knowledge of its derivatives or inflections" (Beglar & Hunt, 1999, p. 147). This assumption is definitely problematic as has been demonstrated in several studies (e.g., Zimmermann & Schmitt, 2002; Ward & Chenjundaeng, 2009; Kremmel & Schmitt, 2016). In addition, Xing and Fulcher (2007) legitimately criticize the word lists on which the VLT versions are based as being out of date. These criticisms suggest that a new measure of vocabulary breadth is needed. Despite this, there is currently no better measure available for the purpose of diagnosing the written receptive word meaning knowledge of learners at different levels. Cameron (2002), for instance, compared secondary school students' performances on the VLT and Meara's (1992) Yes/No Test and found that the VLT was a more useful tool to profile and diagnose learners' vocabulary knowledge. This was the intended use of the VLT in the first place, as will be discussed in the next section.

Pedagogical Implications

Originally designed as a diagnostic tool for teachers (Nation, 1983, 1990), the VLT has now come to be used as a widely employed instrument amongst teachers and researchers alike to provide an estimate of vocabulary breadth of L2 language learners (Read, 1988; Cobb, 1997; Schmitt & Meara, 1997; Laufer & Paribakht, 1998; Shiotsu & Weir, 2007). However, this is not the intended use of this instrument. Rather than giving an overall estimate of a learner's vocabulary size, the VLT allows for profiling word knowledge at particular levels. This means that scores should not simply be added up. Instead, the four (or five) level sections stand alone and can be administered as the teaching context demands. The sections and levels must be chosen appropriately for the pedagogic purpose and context. Administering the 10,000 level to beginner learners is time poorly spent as it will not yield a lot of useful information. Likewise, very proficient or advanced students do not necessarily have to take the 2,000 level if the teacher thinks this level is not causing lexical problems for the learners.

Having two parallel test versions enables teachers to administer two similar tests to the same *group* on two occasions without any memory effect confounding results. However, the two parallel tests were not found to be equivalent enough to be used to measure the learning gains of any *individual learner*, and for this purpose, the same version should be used twice. It also needs to be stressed that the VLT is not a test to be administered on a regular (weekly or monthly) basis. A substantial amount of time and learning should pass between administrations as it is not fine-grained enough as a measure to gauge the small incremental gains and losses that characterize vocabulary development.

Scores on the VLT should be interpreted carefully. As mentioned above, they are not a total vocabulary size estimate. Also, the test measures only a basic level of receptive form-meaning link knowledge. The assumption is often made by end-users that this should enable reading, but that assumption is debatable given that the item format taps into only very basic form-meaning link knowledge. Unfortunately, to date, no research has been carried out to determine whether words correctly answered on the VLT can in fact be understood when reading. Furthermore, scores are not an indication of how well someone can use a particular word in their language production. The test does also not allow for any valid inferences about a person's depth of vocabulary knowledge in terms of multiple meanings of a word, or other aspects of word knowledge such as collocations or register (Nation, 2001).

Its design rationale of sampling items according to particular frequency levels in the VLT, does however enable score interpretations that can be linked to learners' ability to do various things with language. Although there is still a dearth of research on the lexical requirements for language production, recent research has developed some of the figures for oral and written reception. Van Zeeland (2010) found that around 2,000–3,000 word families are needed for conversational listening. Adolphs and Schmitt (2003) concluded that about the same amount of word families enables engagement in basic daily conversation. Webb and Rodgers

(2009) demonstrated that learners require about 3,000 word families to watch and largely understand movies and television programs, thereby confirming the importance of the 2,000 and the 3,000 level of the VLT. The figures required for written reception, that is, reading, which is the primary focus of the VLT, are higher, with Nation (2006) calculating that 8,000–9,000 word families are necessary to be able to read widely. This level can be considered covered by the VLT's 10,000 level.

These vocabulary size requirements led Schmitt and Schmitt (2014) to argue for the teaching, and therefore testing, of mid-frequency vocabulary of between 3,000 and 9,000 word families. While 3,000 words might be enough to arrive at reasonable initial comprehension when accessing authentic listening, movie viewing, and reading texts, knowledge of these additional mid-frequency word families would certainly make any of these experiences less strenuous and thus more enjoyable (for an in-depth discussion of this issue, see Schmitt & Schmitt, 2014). New measurement instruments which include more mid-frequency levels are therefore very desirable, but these alternative tools have yet to be designed and demonstrated to yield similarly valid and reliable results.

Until new properly-validated tests are developed, the VLT with its four levels appears to match these critical frequency levels closely enough to still be a useful tool for teachers in diagnosing lexical profiles and helping to tailor not only vocabulary support, but also teaching materials to the learners' levels and needs. Used in this way, the VLT can assist in selecting appropriate reading materials, for instance the appropriate level of graded reader. Such informed selection and development of reading materials texts with suitable coverage is crucial as learners will simply give up when too many words are unknown (Waring & Nation, 2004; Nation, 2006).

SEE ALSO: Levels or Stages of Word Knowledge; Lexical Approach; Teaching Aids and Materials

References

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, *24*(4), 425–38.

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing*, *16*(2), 131–62.

Cameron, L. (2002). Measuring vocabulary size in English as an additional language. *Language Teaching Research*, *6*(2), 145–73.

Cobb, T. (1997). Is there any measurable learning from hands-on concordancing? *System*, *25*, 201–315.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213–38.

Huhta, A., Alderson, J. C., Nieminen L., & Ullakonoja, R. (2011). *Diagnosing reading in L2—predictors and vocabulary profiles.* Paper presented at the ACTFL CEFR Alignment Conference, Provo, UT.

Kamimoto, T. (2008). *Guessing and vocabulary tests: Looking at the vocabulary levels test*. Paper presented at the 41 Annual BAAL Conference, Swansea, UK.

Kremmel, B., & Schmitt, N. (2016). Interpreting vocabulary test scores: What do various item formats tell us about learners' ability to employ words? *Language Assessment Quarterly*, *13*(4), 377–92.

Laufer, B., & Paribakht, T. S. (1998). The relationship between passive and active vocabularies: effects of language learning context. *Language Learning*, *48*, 365–91.

Meara, P. (1992). *EFL vocabulary tests*. Swansea, Wales: Lognostics.

Meara, P. (1994). The complexities of simple vocabulary tests. In F. G. Brinkman, J. A. van der Schee, & M. C. Schouten-vanParreren (Eds.), *Curriculum research: Different disciplines and common goals* (pp. 15–28). Amsterdam, Netherlands: Vrije Universiteit.

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

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

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Boston, MA: Heinle and Heinle.

Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge, England: Cambridge University Press.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, *63*, 59–82.

Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal*, *19*(2), 12–25.

Schmitt, N. (2000). *Vocabulary in Language Teaching*. Cambridge, England: Cambridge University Press.

Schmitt, N. (2010). *Researching Vocabulary: A Vocabulary Research Manual*. Basingstoke: Palgrave Macmillan.

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.

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*(4), 484–503.

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.

Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing*, *24*(1), 99–128.

Stewart, J., & White, D. (2011). Estimating guessing effects on the vocabulary levels test for differing degrees of word knowledge. *TESOL Quarterly*, *45*(2), 370–80.

Van Zeeland, H. (2010). *Lexical coverage and L2 listening comprehension: How much does vocabulary knowledge contribute to understanding spoken language?* (Unpublished MA dissertation), University of Nottingham.

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

Webb, S., & Rodgers, M. P. H. (2009). The lexical coverage of movies. *Applied Linguistics*, *30*, 407–42.

Waring, R., & Nation, I. S. P. (2004). Second language reading and incidental vocabulary learning. *Angles on the English Speaking World*, *4*, 97–110.

Xing, P., & Fulcher, G. (2007). Reliability assessment for two versions of vocabulary levels tests. *System*, *35*(2), 182–91.

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication*, *3*(2), 215–29.

Suggested Reading

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.