# The Creation of a New Vocabulary Levels Test

Stuart McLean[1] and Brandon Kramer[2]
stuart93@me.com
*1. Kansai University Graduate School*
*2. Momoyama Gakuin University*

## Abstract

This paper describes a new vocabulary levels test (NVLT) and the process by which it was written, piloted, and edited. The most commonly used Vocabulary Levels Test (VLT) (Nation, 1983, 1990; Schmitt, Schmitt, & Clapham, 2001), is limited by a few important factors: a) it does not contain a section which tests the first 1,000-word frequency level; b) the VLT was created from dated frequency lists which are not as representative as newer and larger corpora; and c) the VLT item format is problematic in that it does not support item independence (Culligan, 2015; Kamimoto, 2014) and requires time for some students to understand the directions. To address these issues, the NVLT was created, which can be used by teachers and researchers alike for both pedagogical and research-related purposes.

Keywords: vocabulary, assessment, levels, vocabulary levels test, vocabulary size

The purpose of this article is to provide a clear description of a new vocabulary levels test (NVLT) to assist teachers and researchers in its use. The NVLT was created as a parallel written receptive form of the Listening Vocabulary Levels Test (LVLT) (McLean, Kramer, & Beglar, 2015) and its creation therefore followed similar guidelines (see www.lvlt.info).

Vocabulary tests are often conceptualized as measuring either receptive or productive vocabulary knowledge, estimating either the total number of vocabulary items known (size tests) or mastery of vocabulary at certain frequencies of occurrence within a given corpus (levels tests). This paper introduces a new vocabulary levels test (NVLT), a receptive test of the most frequent 5,000 word families in Nation's (2012) British National Corpus / Corpus of Contemporary American English (BNC/COCA) word list. As the purposes and score interpretations of size and levels tests are often muddled within published research, the differences between the two types will be explained before describing the creation and intended interpretation of NVLT scores.

### Measuring vocabulary size and interpreting vocabulary size test scores

Vocabulary size tests are intended to estimate the total number of words a learner knows. This estimate can be useful when comparing groups of learners, measuring long-term vocabulary growth, or providing "one kind of goal for learners of English as a second or foreign language" (Nation, 2013, p. 522). The Vocabulary Size Test (VST) (Nation & Beglar, 2007), for example, is a measure of written receptive word knowledge based on word family frequency estimates derived from the spoken subsection of the BNC (Nation, 2006). Each item on the VST presents the target word first in isolation followed by a non-defining context sentence, with four answer-choices presented in either English or in the learners' L1. Results of the VST among samples with a wide range in ability have shown that the test is able to reliably distinguish between learners of different vocabulary proficiency, either using the monolingual version (Beglar, 2010) or the various bilingual variants (Elgort, 2013; Karami, 2012; Nguyen & Nation, 2011).

Despite the VST's utility in separating students as a general measure of written receptive vocabulary knowledge *breadth*, inferences based on these results should be made with caution. For example, one of the stated interpretations of the VST is as an approximate estimate of known vocabulary. As the test samples 10 words each from the most frequent 1,000-word frequency bands (up to the 14th or 20th band depending on the version), "a test taker's score needs to be multiplied by 100 to get their total vocabulary size" (Nation, 2013, p. 525). A score of 30 out of 140, for example, would produce a size estimate of

3,000 known word families. While this score interpretation seems straightforward, it carries with it two assumptions which must be addressed: a) the target words on the VST are representative of the frequency bands which they were sampled from, so that each target word can be considered to represent 100 others, and b) correctly answering an item implies the written receptive knowledge of that target word. The first assumption, that the target words on the VST are representative of the frequency bands which they were sampled from, can be sufficiently assumed because the words were randomly sampled according to Nation and Beglar (2007). The second assumption, however, is a bit more problematic as the item format utilizes a 4-choice multiple-choice format, implying a 25% chance that the item would be correctly answered even if the examinee has absolutely no knowledge of the target word. While Nation (2012) recommends that all participants complete the entire 14,000-word version of the VST, McLean, Kramer, and Stewart (2015) showed that most correct answers for low proficiency students at the lowest frequency bands could be attributed to chance rather than lexical knowledge.

In order to increase the accuracy of the VST results, Beglar (2010), Elgort (2013), and McLean, Kramer, and Stewart (2015) recommend that students only take the test two levels above their ability. While this would reduce the previously mentioned score inflation due to mismatched items, the resultant score would not hold much pedagogical value. While some suggest that a VST score can be used to assign reading materials (Nation, 2013; Nguyen & Nation, 2011), this claim ignores the properties of the construct being measured (vocabulary *breadth*) as well as findings which argue that comprehension of reading materials require learners to know at least 95% of the words within the materials (e.g. Hsueh-chao & Nation, 2000; Laufer, 1989; van Zeeland & Schmitt, 2013). This is because while a vocabulary size score can give a rough estimate of the amount of words known, it does not imply knowledge of all vocabulary within that size estimate. For example, McLean, Hogg, and Kramer (2014) reported that the mean vocabulary size of Japanese university students ($N$ = 3,427) was 3,396 word families ($SD$ = 1,268) using the VST. These same learners, however, could not be said to have knowledge of the most frequent 3,396 word families, as all but the most able students had gaps in their knowledge of items from the first 1,000 words of English and all students failed to correctly answer multiple-choice items at the second and third 1,000-word bands.

Similar gaps have been found with the first and second 1,000-word frequency bands by Beglar (2010), Elgort (2013), Karami (2012), and Nguyen & Nation (2011). In order to measure knowledge of the most frequent vocabulary levels, a test made for that purpose is more appropriate.

## Measuring knowledge of vocabulary levels and interpreting VLT scores

While the VST may be an appropriate instrument for separating students with a wide range of proficiencies, a more pedagogically useful measure of lexical knowledge is a test designed to measure the degree of mastery of the most frequent words of English. The most well-known of such tests, the Vocabulary Levels Test (VLT) (Nation, 1990; Schmitt, et al., 2001) was designed to provide richer information about learners' knowledge of the second, third, fifth, and tenth 1,000-word frequency bands, as well as Coxhead's (2000) Academic Word List (AWL). The primary purpose of a levels test such as this is to estimate learners' mastery of the most frequent vocabulary in the hope of assigning appropriate learning materials. For example, Nation (2013) states that meaning-focused reading input, which would include activities such as extensive reading and many kinds of task-based instruction, requires instructional materials to be written at a level with 95% known vocabulary. The test scores and their interpretations reflect this purpose, usually represented as a score out of 30 items for each level of the test, with mastery being a high proportion of correct answers at that level. Teachers can then use these results to help students focus on the most frequent unknown words until mastery is achieved.

## Limitations of previous written vocabulary levels tests

While many have found the VLT (Nation, 1983, 1990; Schmitt, et al., 2001) useful in both pedagogy and research, Webb and Sasao (2013) identified a number of issues which this paper and the NVLT described within attempt to address.

### *Previous VLT content*

The first limitation of the previous versions of the VLT is the absence of a section testing knowledge of the first 1,000-word frequency level, considered to be of the greatest value to learners because of the impact high frequency words have on comprehension. While the word families within the first 1,000-word frequency level account for 78% of the BNC corpus, the words from the next most frequent 1,000 word families account for only 8.1% (Nation, 2013).

Second, previous versions of the VLT sampled target words and distractors from outdated frequency lists. The first and second 1,000-word frequency levels used words from West's (1953) General Service List (GSL), and the 3,000, 5,000, and 10,000 word-frequency bands sampled words from lists constructed from Thorndike and Lorge (1944) and Kučera and Francis's (1967) frequency criteria. These lists represent the best effort to represent the language at the time they were made, but languages, and the vocabulary within them, are known to drift over time. In addition, advances in technology over the past few decades have allowed for improved corpus building, allowing researchers to collect much larger and more representative samples, which can be analyzed much more efficiently and accurately than the lists used to construct the VLT, allowing teachers to measure knowledge of vocabulary which would be considered much more appropriate for language learners today.

### *Previous VLT format*

The previous VLT format (see Figure 1 for an example item cluster), which presents target items with distractors of the same vocabulary level, is problematic for several reasons: a) a lack of item independence, b) the relative inaccuracy of the format when compared with a standard four-choice item, c) student difficulty understanding the format, and d) difficulty adapting the tests to other testing mediums or base corpora.

| | |
|---|---|
| 1 business | |
| 2 clock | _____ part of a house |
| 3 horse | _____ animal with four legs |
| 4 pencil | _____ something used for writing |
| 5 shoe | |
| 6 wall | |

Figure 1. *Example of the VLT format.*

An assumption of test item analyses, whether within classical testing theory or item response theory (IRT), is that the items demonstrate what is called *item independence*. This means that the responses to different test items are not dependent on each other, meaning that they need to measure distinct aspects of knowledge. The VLT format (see Figure 1) displays six answer choices on the left, to be matched with the three target word definitions on the right. As students answer the three items, the number of available answer choices decreases, allowing them to answer more easily. Because of this, during their validation of VLT data, Beglar and Hunt (1999, p. 154) stated that "it has not been shown that the assumption of item independence holds true given this test format", a concern supported by Culligan (2015). Kamimoto (2014), looking into this issue specifically, concluded that the VLT format interacts with examinees'

knowledge of target items and causes local item dependence to various degrees and that this violation of item independence "comes from the test format" (p. 56).

Recently, Kremmel (2015) investigated the behavior of the different test formats in relation to qualitative interviews where the participants demonstrated knowledge of the target words. While both the item cluster VLT format and the standard multiple-choice format of the VST performed reasonably well, Kremmel found that the VLT format was slightly less representative of the participants' actual knowledge. This evidence suggests that the multiple-choice format more accurately measures vocabulary knowledge than the old levels test format, relative to the criterion of recall of meaning.

Previous use and piloting of the VLT format suggested that examinees may hesitate in answering VLT items and find its format problematic. The tests were piloted in a low English proficiency high school, and much time was necessary in order to carefully explain the testing procedure and allow the students to work through practice problems. In contrast, the standard multiple-choice format was immediately understood by the examinees, which facilitated a quicker administration of the test.

Finally, a standard multiple-choice format is also more easily adapted to online tests using widely available online testing software such as Survey Monkey <surveymonkey.com> or Moodle <moodle.org>, allowing teachers, researchers, and policy-makers to quickly administer and analyze tests or surveys with a large number of participants. A related limitation of this format it is that the distractors are not as easily edited as those within a standard multiple-choice item, as all distractors have to be considered in relation to the three target meanings. This would be particularly troublesome, for example, if a researcher tried to reorder the items to reflect a different wordlist which orders words differently, a problem further exacerbated if the lists utilize different word counting units.

## The New Vocabulary Levels Test

In order to address the limitations stated above and provide an instrument with greater pedagogical utility, the authors created a new vocabulary levels test (NVLT). This NVLT is intended as a diagnostic and achievement instrument for pedagogical or research purposes, measuring knowledge of English lexis from the first five 1,000-word frequency levels of the BNC and the Academic Word List (AWL) (Coxhead, 2000). The test consists of five 24-item levels which together measure knowledge of the most frequent 5,000 word families, in addition to a thirty-item section which measures knowledge of the AWL. The entire 150-item test can be completed in 30 minutes; however, depending on the specific needs of researchers or teachers specific test sections can be administered in isolation.

### *NVLT format*

The NVLT utilizes the multiple-choice format which provides multiple benefits: a) manipulation of distractor difficulty; b) efficient and reliable electronic marking; c) easily conducted item analyses; and d) item independence, a prerequisite for test analysis. Each item consists of four answer choices, from which examinees must select the word or phrase with the closest meaning to the target word. An example item is shown in Figure 2.

1. time: They have a lot of **time**.
a. money
b. food
c. hours
d. friends

*Figure 2*. An example item from the NVLT.

The piloted and revised test instructions (see Appendix A) are presently available in English and Japanese, with plans for additional languages in the future. When possible, to ensure that test instructions are understood, they should be given to examinees in their first language. To reduce the effects of guessing, the instructions state that if examinees have no knowledge of the correct answer, they should skip the question. However, if examinees feel that they may know the word, they should answer. The instructions also include two example questions to encourage understanding of the test format. Teachers and researchers should use the instructions they feel most appropriately meet their needs, while remembering that altering the instructions of the test may alter how items function.

### *The source of target vocabulary*

The target words of the NVLT come from Nation's (2012) British National Corpus (BNC)/Corpus of Contemporary American English (COCA) word lists. The first and second 1,000-word family lists of the BNC/COCA were derived from a 10 million token corpus that consists of 6 million tokens from spoken British and American English. The corpus provides a list of high frequency words suitable for teaching and course design, and is a separate corpus than the one used to make the third to twenty-fifth 1,000-word family bands (Nation, 2012). The lists for the third 1,000-word family and above were created from BNC/COCA rankings once word families from the first 2,000 words of the BNC/COCA were removed. The BNC/COCA word lists include both British and North American varieties of English and are partly based on a spoken corpus, providing a strong basis for a monolingual vocabulary test (Nation, 2012). As Webb and Sasao (2013) stated, "the new BNC/COCA lists should be representative of current English and provide a far better indication of the vocabulary being used by native speakers today than the lists used for the creation of the earlier versions of the VLT" (p. 267).

The NVLT utilizes the word family unit because a) it was the unit utilized during the creation of the twenty-five 1,000-word BNC/COCA frequency lists (available with the Range software program, Heatley & Nation, 2015), b) even low proficiency learners have some control of word-building devices and they can perceive both a formal and semantic relationship between regularly affixed members of a word family (Nation & Beglar, 2007), c) it is consistent with the parallel LVLT and previous levels tests allowing for better comparison, and d) there is evidence that the word family is a psychologically real unit (Bertram, Baayen, & Schreuder, 2000; Bertram, Laine, & Virkkala, 2000; Nagy, Anderson, Schommer, Scott, & Stallman, 1989).

If given in its entirety the NVLT can measure knowledge of the first five 1,000-word frequency levels of the BNC/COCA and the AWL, which provides adequate coverage for numerous reading genres. As Webb and Sasao (2013) stated, "mastery of the 5,000 word level may be challenging for all but advanced learners, so assessing knowledge at the five most frequent levels may represent the greatest range in vocabulary learning for the majority of L2 learners" (p. 266).

### *Test creation*

The items making up the first five 1,000-word frequency levels of the NVLT were created through a process of retrofit and redesign of previous Vocabulary Size Test (VST) items (Nation & Beglar, 2007). The previous validation of the use of the VST items with Japanese university students in an EFL context (Beglar, 2010) suggested their appropriateness to the NVLT which was piloted with a similar group. Item specifications (see Appendix B) were reverse engineered from previous test descriptions (e.g. Nation & Beglar, 2007) and specification-driven test assembly was implemented in line with Fulcher and Davidson (2007) when retrofitting items from three monolingual VST versions. Two VST versions were downloaded from <victoria.ac.nz/lals/about/staff/paul-nation> while the third version was obtained through personal correspondence with I.S.P. Nation. Items were re-assigned to their appropriate

BNC/COCA levels. For example, *period* and *basis* were relocated from the first 1,000-word level to the second 1,000-word level and items such as *nil*, present in the second 1,000-word frequency level of the VST, are not present in the NVLT as they do not occur in the first five 1,000-word levels of the BNC/COCA lists.

To ensure that the test is not conflating the construct of L2 contextual inferencing with vocabulary knowledge, the context sentences for each item were piloted using pseudowords in place of the target words. If the participants were then able to identify the correct answer without seeing the target word, the context sentence was edited as necessary.

The NVLT includes the AWL for three reasons: a) the importance of accessing AWL vocabulary knowledge because of the prominence of academic English programs; b) 10% coverage of tokens in academic texts is provided by the AWL (Coxhead, 2000); and c) previous tests measuring knowledge of the AWL have relied on the problematic VLT format.

AWL items were also created using the item specifications listed in Appendix B. The AWL is divided into nine 60-word and one 30-word sublists according to word frequency (Coxhead, 2000). Three target words were chosen from each of the first nine sublists and two from the tenth using a random number generator, and the final item was chosen at random from the entire AWL to ensure an even distribution of items. The final target word within each test item was the headword of the AWL word family (as listed in Coxhead, 2000). After each target word was chosen, distractor choices were randomly selected from the same sublist as the target word until the desired part of speech was obtained. If a suitable distractor could not be found in the same sublist, the process was repeated one sublist lower (i.e., the next higher frequency sublist).

Piloting was conducted to ensure that all distractors were plausible options. Then a generic sentence providing context without assisting the selection of the correct answer was written for each selected target item. Concordancer output from <www.lextutor.ca/conc/eng/> using the BNC/COCA corpus was consulted for authentic examples when the target word had numerous uses or meanings. When a sentence did not fit all of the distractors, the non-conforming distractor was replaced with randomly chosen words until all were found to fit the necessary criteria. Finally, each example sentence was checked to ensure that words from the first 1,000-word frequency level were used; however, a very limited number of words from the second 1,000 words of English were included, which were not found to be a problem in pilot testing. Repeated piloting of a small number of items continued until all significant problems were resolved.

### *Interpretability*

Test interpretability is the degree to which qualitative meaning can be assigned to the quantitative measures produced by a test instrument (Medical Outcomes Trust Scientific Advisory Committee, 1995), and it is important for test creators to explicitly state how the test scores can be interpreted. The NVLT is intended as a test that measures an examinee's knowledge of the written form-meaning link of decontextualized vocabulary frequency bands. As a result, NVLT test scores should not be used to make statements about an examinee's productive vocabulary knowledge (see Laufer & Nation, 1999) or receptive aural vocabulary knowledge (see McLean, Kramer, & Beglar, 2015). It is recommended that the NVLT be utilized as a diagnostic, formative, or summative instrument, and that researchers and teachers use the 1,000-word frequency bands of the test that are appropriate for their needs. It is not recommended that the number of items per 1,000-word frequency level be reduced without careful IRT analysis.

While further research and testing is needed to empirically show the NVLT's utility in a variety of contexts, we can hypothesize potential uses for teachers and researchers. One example of an appropriate use of the

NVLT would be to assess learners' readiness for a particular course of study or the appropriateness of materials for learners. Instructors could first estimate the written vocabulary load of instructional materials or a single text. Given that research has shown that 98% coverage is ideal for easily comprehending written material (Hsueh-chao & Nation, 2000), the NVLT can be used to estimate learners' knowledge of lexis at particular word-frequency levels to determine whether they have the necessary lexical knowledge to comprehend course materials. For instance, learners who correctly answer at least 47-48 of the 48 items from the 1,000 and 2,000 word-frequency levels and half of the items from the 3,000 word-frequency levels on the NVLT would be deemed to have sufficient lexical knowledge to comprehend texts consisting of the most frequent 2,000 English word families. It should be remembered that this test is based on BNC/COCA word family lists. Thus, using the NVLT to assign level appropriate materials written based on different wordlists, and especially wordlists which use the lemma counting unit, is not recommended.

The NVLT could also be used to diagnose learners' vocabulary knowledge at the beginning of a course of study, estimate achievement throughout the course of study (i.e., formative assessment), and measure the knowledge gained upon completion of a course (i.e., summative achievement). For instance, if the goal of a beginner level course is to acquire knowledge of the 2,000 most frequent words of English, the threshold for mastering a single 1,000-word level should be at least 23 out of 24 correct items. Importantly, for higher frequency bands the necessity for a high mastery threshold is crucial, as any language user will commonly meet the highest frequency words when using the target language. This strict threshold is further supported by the mixed-methods validation of the aural version of this test (McLean, Kramer, & Beglar, 2015), which found that test-takers were more likely to correctly guess items that they did not know than to miss items that they knew. Similarly, mastery of the most frequent academic vocabulary should be defined as correctly answering 29 or more of the 30 AWL items.

## Conclusion

The NVLT is a test that measures examinees' written receptive knowledge of the most frequent vocabulary frequency bands. The NVLT possesses four advantages over versions of the previous VLT: a) it measures vocabulary knowledge of each of the first five 1,000-word frequency bands; b) it measures vocabulary knowledge based on the more comprehensive and recent BNC/COCA; c) it utilizes a multiple-choice format facilitating item independence; and d) it has a parallel aural vocabulary levels test, the LVLT. It is recommended that the NVLT be used as a diagnostic, formative, or summative instrument, and that researchers and teachers utilize the 1,000-word frequency bands of the test that are appropriate for their needs. The test form is freely available and can be downloaded from <lvlt.info> or by contacting the authors.

## References

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

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

Bertram, R., Baayen, R. H., & Schreuder, R. (2000). Effects of family size for complex words. *Journal of Memory and Language*, *42*(3), 390-405. doi: 10.1006/jmla.1999.2681

Bertram, R., Laine, M., & Virkkala, M. M. (2000). The role of derivational morphology in vocabulary acquisition: Get by with a little help from my morpheme friends. *Scandinavian Journal of Psychology*, *41*(4), 287-296. doi: 10.1111/1467-9450.00201

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

Culligan, B. (2015). A comparison of three test formats to assess word difficulty. *Language Testing, 32*(4), 503-520. doi: 10.1177/0265532215572268

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing, 30*(2), 253-272. doi: 10.1177/0265532212459028

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. New York: Routledge.

Heatley, A., & Nation, I. S. P. (2015). Range. Retrieved from http://www.victoria.ac.nz/lals/about/staff/publications/BNC_COCA_25000.zip

Hsueh-chao, M. H., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403-430. Retrieved from http://nflrc.hawaii.edu/rfl/PastIssues/rfl131hsuehchao.pdf

Kamimoto, T. (2014). Local item dependence on the Vocabulary Levels Test revisited. *Vocabulary Learning and Instruction, 3*(2), 56-68. doi: 10.7820/vli.v03.2.kamimoto

Karami, H. (2012). The development and validation of a bilingual version of the Vocabulary Size Test. *RELC Journal, 43*(1), 53-67. doi: 10.1177/0033688212439359

Kremmel, B. (2015). *The more, the merrier? Issues in measuring vocabulary size*. Paper presented at the LTRC 2015: The Language Testing Research Colloquium, Toronto.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English.* Providence, R.I.: Brown University Press.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special Language: From Humans Thinking to Thinking Machines*. Clevedon: Multilingual Matters.

Laufer, B., & Nation, I. S. P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing, 16*(1), 33-51. doi: 10.1191/026553299672614616

McLean, S., Hogg, N., & Kramer, B. (2014). Estimations of Japanese university learners' English vocabulary sizes using the Vocabulary Size Test. *Vocabulary Learning and Instruction, 3*(2), 47-55. doi: 10.7820/vli.v03.2.mclean.et.al

McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research, 19*(6), 741-760. doi: 10.1177/1362168814567889

McLean, S., Kramer, B., & Stewart, J. (2015). An empirical examination of the effect of guessing on vocabulary size test scores. *Vocabulary Learning and Instruction, 4*(1), 26-35. doi: 10.7820/vli.v04.1.mclean.et.al

Medical Outcomes Trust Scientific Advisory Committee. (1995). Instrument review criteria. *Medical Outcomes Trust Bulletin, 3*(4), 1-4.

Nagy, W., Anderson, R. C., Schommer, M., Scott, J. A., & Stallman, A. C. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly, 24*(3), 262-282. doi: 10.2307/747770

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines, 5*(1), 12-25. Retrieved from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/1983-Testing-and-teaching.pdf

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Rowley, Mass.: Newbury House.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*(1), 59-82. doi: 10.3138/cmlr.63.1.59

Nation, I. S. P. (2012). The BNC/COCA word family lists. Retrieved 17 September, 2012, from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Information-on-the-BNC_COCA-word-family-lists.pdf

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge: Cambridge University Press.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9-13. Retrieved from http://jalt-publications.org/files/pdf/the_language_teacher/07_2007tlt.pdf

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. doi: 10.1177/0033688210390264

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. doi: 10.1177/026553220101800103

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 Words*. New York: Teachers College Press, Columbia University.

van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics, 34*(4), 457-479. doi: 10.1093/applin/ams074

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

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

# Appendix A

## Translation of NVLT instructions

This is a vocabulary test.

Please select the option a, b, c or d which has the closest meaning to the word in **bold**.

**Example question**
**see**: They **saw** it.
a. cut
b. waited for
c. looked at The correct answer is **c**.
d. started

If you have no idea of the answer at all, please do not answer the question and move on to the next question.

However, if you think there is a chance that you may know the word, please try to answer.

Let's begin.

## New Vocabulary Levels Test: 説明

これは単語力テストです。

太字になっている英語の意味に最も合う選択肢を a～d から選んでください。

**問題例**

**see**: They **saw** it.

a. cut

b. waited for

c. looked at 正解は **c** です。

d. started

答えが全く分からない場合は、空白にしておいてください。

しかし、わかる可能性があると思ったら、どうぞ挑戦してみてください。

では、始めましょう。

# Appendix B

## Specifications for New Items

*Example Item:*

school: This is a big **school**.
a. where money is kept
b. sea animal
c. place for learning
d. where people live

*Overall*

- The target word is presented in isolation and in bold within a context sentence
- The answer key should be randomly generated
- Avoid gender-biased language and have balanced gender representation

*Target words*

- Written in citation form
- From frequency list based on established corpus (BNC/COCA)
- Random sampling of words from each word-frequency level

*Context sentence*

- Context sentences in the first two 1,000-word levels should be written using vocabulary within the first 1,000-word level whenever possible
- Context sentences in the third 1,000-word level and above should be written using vocabulary within the first two 1,000-word levels whenever possible
- In cases where the part of speech is ambiguous, the most common form should be used based on frequency data
- The accompanying sentence should be as contextualized as possible without giving hints to the meaning of the target word

*Distractors*

- Core meanings of distractors should be of similar word frequency and difficulty level as the target word
- Distractors for items in the first two 1,000-word levels should be written using vocabulary within the first 1,000-word level whenever possible
- Distractors in the third 1,000-word level and above should be written using vocabulary within the first two 1,000-word levels whenever possible
- To as great a degree as possible, all distractors should be equally plausible in the context sentence