The updated Vocabulary Levels Test

Developing and validating two new forms of the VLT

Stuart Webb, Yosuke Sasao and Oliver Ballance
University of Western Ontario / Kyoto University / Victoria University of Wellington

The Vocabulary Levels Test (Nation, 1983; Schmitt, Schmitt, & Clapham, 2001) indicates the word frequency level that should be used to select words for learning. The present study involves the development and validation of two new forms of the test. The new forms consist of five levels measuring knowledge of vocabulary at the 1000, 2000, 3000, 4000, and 5000 levels. Items for the tests were sourced from Nation's (2012) BNC/COCA word lists. The research involved first identifying quality items using the data from 1,463 test takers to create two equivalent forms, and then evaluating the forms with the data from a further 250 test takers. This study also makes an initial attempt to validate the new forms using Messick's (1989, 1995) validity framework.

Introduction

The Vocabulary Levels Test (VLT) is perhaps the most widely used measure of L2 lexical knowledge (Read, 2000). It was originally developed by Nation (1983) and then updated by Schmitt, Schmitt, & Clapham (2001) as a means to determine the extent to which test takers could recognize the form-meaning connections of words at four word frequency levels (2000, 3000, 5000, 10000) and an academic vocabulary level. The test can be done as a whole with students completing all levels, or it can be done with only individual levels. For example, it is probably only necessary to administer the 2000 word level to beginners since they are unlikely to have mastered any of the subsequent levels. The greatest value of the VLT is that it indicates at which word frequency level students should focus their learning.

The VLT employs a matching format in which the participants are presented with 30 questions per level. The words are presented in 10 clusters of six words (three keys and three distractors) and three definitions at each level. The test taker's job is to write the correct item numbers beside their corresponding definitions.

ITL – International Journal of Applied Linguistics 168:1 (2017), **33–69**. DOI 10.1075/itl.168.1.02web
ISSN 0019–0829 / E-ISSN 1783–1490

For each correct response in a cluster, the participant receives a point, so the maximum score at each level is 30. When scoring the test, the scores for the individual levels are most important because these scores reveal where subsequent vocabulary learning should be focused. In contrast, the overall score has little meaning. The items in 5 of the 10 clusters are made up of nouns. The items in 3 of the clusters are verbs, and the items in 2 of the clusters are adjectives. The proportion of nouns, verbs, and adjectives is representative of their proportional occurrence in English although it should be noted that this may vary within frequency bands. Figure 1 shows an example of a noun cluster at the 3000 level in one of Schmitt, Schmitt, & Clapham's (2001) versions of the VLT.

1	bull	________	formal and serious manner
2	champion	________	winner of a sporting event
3	dignity	________	building where valuable objects are shown
4	hell		
5	museum		
6	solution		

Figure 1. Noun cluster from Schmitt, Schmitt, & Clapham's (2001) VLT

Schmitt, Schmitt, & Clapham's (2001) new forms of the test improved on the earlier ones by increasing the number of items per level from 18 to 30 to improve reliability, and selecting academic words from Coxhead's (2000) Academic Word List rather than the original source: Xue & Nation's (1984) University Word List. While these changes greatly improved upon the original version, Schmitt, Schmitt, & Clapham's VLT still had two limitations (Webb & Sasao, 2013). First, items within the word frequency levels were derived from texts from the 1930s and 1940s, and therefore might not reflect current vocabulary. Second, the earlier forms of the VLT did not measure knowledge of the most frequent 1000 word families. This is particularly important because the relative value of words has a marked decrease after the most frequent 1000 word families; the most frequent 1000 word families account for as much as 80% of English, while the most frequent 1001 to 2000 word families make up from around 4 to 10% of English. Thus, the most valuable word frequency level to measure is the most frequent 1000 word families because of its importance to understanding English. The aim of the present study was to create new forms of the VLT to overcome these limitations.

Creating new forms of the Vocabulary Levels Test

The new forms of the VLT that were developed in this study followed the principles used to create the earlier versions. The new forms also use a matching format with 10 3-item clusters per level and measure knowledge of the same proportions of nouns, verbs, and adjectives (15, 9, and 6 items per level, respectively) as the earlier versions. However, there were three major changes made to the new versions. First, although the new forms were made up of five levels, the word frequency levels were changed. The five word frequency levels in the new forms were 1000 (the most frequent 1–1000 word families), 2000 (the most frequent 1001–2000 word families), 3000 (the most frequent 2001–3000 word families), 4000 (the most frequent 3001–4000 word families), and 5000 (the most frequent 4001–5000 word families).

The word family rather than the lemma was used as the unit of counting for several reasons. First, the rationale for counting words as families is that if someone knows a form of a word (e.g., accuse or adventure), they might be able to understand an unknown form when it is encountered (e.g., accuser, accusation, accusingly; adventurer, adventurous, misadventures) with relatively little effort. It is important to note that this argument may only hold true of receptive knowledge (understanding a derivation when it is encountered when reading or listening) rather than productive knowledge (producing an unknown derivation for a known word). Second, the earlier versions of the VLT have been found to be effective diagnostic measures of vocabulary knowledge and all of the earlier forms have used the word family as the unit of counting. Thus, based on earlier studies that have used the VLT, there does not appear to be strong grounds to change the unit of counting. (For further detail on why the word family may be a more useful unit of counting for receptive knowledge than the lemma see Nation, 2016). However, it should be noted that there may also be value in using the lemma as the unit of counting for L2 vocabulary (Dang & Webb, 2016; Kremmel, 2016). This likely depends though on purpose and user. Certainly, the lemma might be a better unit of counting for productive knowledge, because research suggests that L2 learners may often lack productive knowledge of word parts (Schmitt & Zimmerman, 2002). However, the VLT is designed to measure receptive knowledge rather than productive knowledge. Similarly, if test takers are at a beginner level, a lack of receptive knowledge of word parts and limited breadth of knowledge may make it difficult to understand the information that is included within words and contexts to decipher unknown derivations. Thus, it may also be useful to create a more fine-tuned test for beginning learners that is designed to measure knowledge of the most frequent lemmas such as those found in Dang & Webb's (2016) Essential Word List.

There were several reasons for the change in levels. First, it is most important to measure knowledge of the most frequent 1000 word families because this frequency level accounts for by far the greatest proportion of spoken and written English. For example, the most frequent 1000 word families account for around 65–85% of spoken and written English, while the 2000 word level only accounts for around 3–10% of English (Webb & Nation, 2017). The reason that the 4000 word level was also included in the new version and the 10000 word level was excluded was that we believe that it is most useful to provide a profile of the most frequent 5000 word families because these are the most important words for learners. Including these five sequenced levels may allow teachers, learners, and researchers to better evaluate vocabulary learning progress than the previous versions that did not include the 1000 and 4000 levels. Knowledge of the 10000 word level may also provide some indication of progress in a learner's lexical development. However, tests that measure knowledge of vocabulary size such as the Vocabulary Size Test (Nation & Beglar, 2007; Coxhead, Nation, & Sim, 2015), V_YesNo (Meara & Miralpeix, 2017) and CATSS: Computer Adaptive Test of Size & Strength) (Laufer & Levitzky-Aviad, 2016) will provide a more accurate measure of this than a VLT that includes a 10000 word level. Moreover, the inclusion of the 10000 word level in the earlier forms may have led some users to incorrectly assume that the VLT is a measure of vocabulary size. However, this is not the case. The VLT may provide a reliable measure of knowledge of particular word frequency levels but to measure vocabulary size a much larger range of word frequency levels needs to be assessed.

A level measuring knowledge of academic vocabulary was also not included in the new versions. The reason for this is that words in Coxhead's (2000) Academic Word List (AWL) vary greatly in their value. Items in the first sublist are encountered in academic text much more than items in the second sublist, and items in that list are encountered more often than items in the third sublist, and so on. Thus, it was believed that it would be more useful to measure knowledge of particular levels of the AWL rather than the AWL as a whole.

The second major change to the new forms of the VLT was that Nation's (2012) British National Corpus/Corpus of Contemporary American English word lists were used as the source of items for the new versions to ensure that the frequency levels of the items better reflected current English. The headwords in each list were ordered from 1 to 1000 and the True Random Number Generator at Random. Org was used to avoid any bias when selecting items. The keys and distractors were quasi-randomly rather than randomly selected from the word lists because function words and adverbs were not chosen. Clusters were created by selecting three keys plus three distractors all with the same part of speech from the same frequency level. The three target words in each cluster needed to have meanings that were sufficiently different to allow test takers to choose the correct response.

The definitions for the keys in each cluster were made up of words from higher frequency levels to reduce the chances that a lack of vocabulary knowledge of the definitions might limit test takers ability to select the correct response. For example, the definitions for items in the 5000 level were made up of words from the 1000–4000 levels, and the definitions for items in the 4000 level consisted of words from the 1000–3000 levels. The one exception to this was for definitions at the 1000 word level. Because there was no higher frequency level to source the words, definitions for the 1000 word level consisted of vocabulary from that level.

The third change to the test was in the presentation of the clusters. The presentation of the matching format was changed to make it more transparent to test takers. In the new format, a grid was provided with the items presented in bold horizontally across the page and the definitions presented vertically down the page. The test takers job was to check the correct item box for each definition. An example from the 1000 word level is shown in Figure 2.

	boy	**rent**	**report**	**size**	**station**	**thing**
how big or small something is						
place buses and trains go to						
young man						

Figure 2. Noun cluster from new form of the 1000 word level

Development of two equivalent forms

This section describes how equivalent forms of the VLT were created. First, a total of 331 clusters (993 items) were created from which we could identify quality clusters. Table 1 shows the details of the 331 clusters. The quality of these 331 clusters was revealed by examining the responses of learners with a wide variety of L1s and cultural backgrounds.

Table 1. Initial development of new forms of the VLT

Level	Noun	Verb	Adjective	Total
1000	34	24	17	75
2000	31	21	12	64
3000	31	29	16	76
4000	30	18	10	58
5000	28	17	13	58
Total	**154**	**109**	**68**	**331**

The new forms of the VLT were written in a web-based format, because it had the following advantages: (1) test-takers could complete the test anywhere and anytime when they had access to the Internet; (2) the web-based format did not allow users to go back to the previous questions nor skip any questions; and (3) between participants randomization of clusters and target words within clusters was used to reduce the potential for an order effect.

Qualtrics <https://www.qualtrics.com/> was used as the testing platform. The system was programmed so that 5 noun, 3 verb, and 2 adjective clusters were randomly chosen for each level from the item bank, resulting in 50 clusters in total. Figure 3 shows an example of a noun cluster in the web-based format. At the end of the test, the test-takers received feedback about their vocabulary levels estimates and brief suggestions for future learning.

	animal	bath	crime	grass	law	shoulder
green leaves that cover the ground	○	○	○	○	○	○
place to wash	○	○	○	○	○	○
top end of your arm	○	○	○	○	○	○

Figure 3. Example of a noun cluster in the web-based format

The next step was to obtain empirical data about the quality of the items and the item difficulties. The participants were recruited by asking language teachers in various countries to have their students take the test. Some of them took it during normal class hours, while others did it outside of the classrooms. A total of 1,463 participants (916 female; 470 male; 77 unspecified) completed the test.[1] Table 2 summarizes the participants' birthplace revealing that the participants had a wide variety of cultural backgrounds. The participants' ages also varied widely (Figure 4).

The data were analyzed using Winsteps 3.92.1 (Linacre, 2016a) based on the Rasch dichotomous model (Rasch, 1960). Rasch analysis was used because it allows test equating, where all the items are put into one item hierarchy. In this study, a concurrent (or one-step) equating was used where all the data were entered into one big array and the items that had not been taken by a test-taker were treated as missing data. Rasch analysis is also helpful in assessing the degree to which the empirical data fit the Rasch model, the mathematical model indicating the probability of success based on the difference between person ability and item difficulty. In this study, we regarded fit statistics of infit *t* and outfit *t* larger than 2

[1] The following respondents were excluded from analysis: test takers (1) who did not complete the test, (2) who spent more than an hour completing it, and (3) whose response patterns were considerably irregular (Rasch outfit $t > 5.0$).

Table 2. Participants' birthplace

Japan	389	Canada	19
Vietnam	148	Brazil	18
Uzbekistan	114	Italy	18
United States	104	Iran	17
Taiwan	89	Spain	13
China	83	Russian Federation	11
Saudi Arabia	73	United Kingdom	11
Czech Republic	29	Iraq	10
Turkey	29	Mexico	10
Australia	21	Other	192
		Unspecified	65

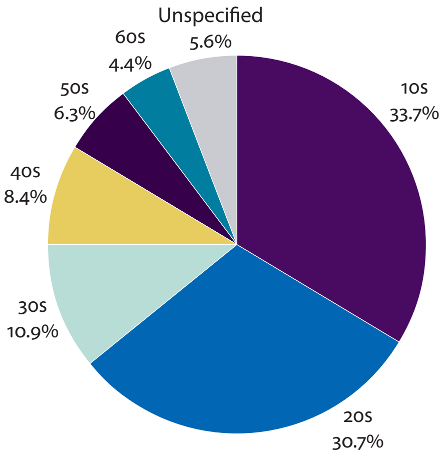


Figure 4. Participants' age distribution

or smaller than −2 as misfit to the Rasch model.[2] A complete cluster was discarded if it had any misfit items. As a result, 234 clusters were found to be acceptable, and 97 clusters were discarded.

[2] Outfit is an unweighted estimate sensitive to unexpected responses by low-ability persons on difficult items or high-ability persons on easy items; infit, on the other hand, is a weighted estimate sensitive to unexpected responses to items targeted on the person (Linacre, 2002). The present research used outfit and infit *t* statistics (instead of unstandardized mean square) as the primary criterion for detecting misfit items, because with large sample size the *t* statistics may identify a greater number of misfit items than mean-square statistics (Karabatsos, 2000; Linacre, 2003; Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008).

Difficulty of each cluster was calculated by averaging the difficulty estimates for the three items in the cluster (Schmitt, Schmitt, & Clapham, 2001). Equivalent forms (Forms A and B) were created based on this "cluster difficulty". The two new forms were created based on the following criteria:

1. Items for the new forms were chosen from the 234 acceptable clusters;
2. Each form has 5 noun, 3 verb, and 2 adjective clusters at each level, resulting in 10 clusters per level and 50 clusters in total;
3. The average cluster difficulty of each level and each part of speech of the new forms approached the original average cluster difficulty of the item bank so that the new forms would reflect the original cluster difficulties; and
4. The new forms had a wide range of cluster and item difficulties.

The following section describes the empirical examination of the quality of the two equivalent forms of the VLT.

Evaluation of the two equivalent forms

This section attempts to answer the following two questions: (1) Are the two new forms of the VLT equivalent? and (2) Do they produce valid and reliable results? More specifically, it discusses the equivalence of the two new forms of the VLT, and then attempts to provide preliminary validity evidence using Messick's (1989, 1995) framework which allows test validation from a wide variety of perspectives.

Instrument

To examine the equivalence of the two new forms, two provisional tests (Provisional Test 1: PT1 and Provisional Test 2: PT2) were created using a common item linking method where the two tests shared 12 clusters in common in order to put all the items into one item hierarchy (Wright & Stone, 1979). For each provisional test, six common clusters[3] that had items with a wide range of difficulty estimates were added from the other test (3 noun clusters, 2 verb clusters, and 1 adjective cluster). Thus, PT1 had all 50 clusters from Form A plus the six common clusters chosen from Form B for a total of 56 clusters. Similarly, PT2 had the 50 clusters from Form B and the six common clusters chosen from Form A, resulting in 56 clusters. In this way, PT1 and PT2 are linked to each other by their 12 common clusters

[3] The term *common cluster* is used to refer to a set of three items in a cluster that were used for common item linking.

and analysis of the results for the two tests may reveal the degree of equivalence between Form A and Form B.

The provisional tests were written in the web-based format using Qualtrics. The two tests were randomly assigned to test takers when they clicked on the URL to complete the test.

Participants

Data were collected from a total of 250 participants (51 male, 196 female, 3 unspecified) learning English in three different countries (Japan, Spain and China). In Japan, 148 university students learning English as a foreign language participated in the study. Their ages ranged between 18 and 21 with an average age of 18.3. They majored in engineering, economics, and education, and their language proficiency level as indicated by self-reported TOEIC® (Test of English for International Communication) Listening & Reading Test scores was $M = 586.1$, $S.D. = 194.4$. They were supervised while they were taking the test. In Spain ($N = 62$) and China ($N = 40$), data were collected from students learning English as a foreign language. Their ages ranged between 19 and 45 with the average being 22.9. Their self-reported CEF-R (Common European Framework of Reference for Languages) levels were C2 (3.2%), C1 (9.7%), B2 (22.6%), B1 (61.3%), A2 (3.2%), and A1 (0%).

Equivalence of the two forms

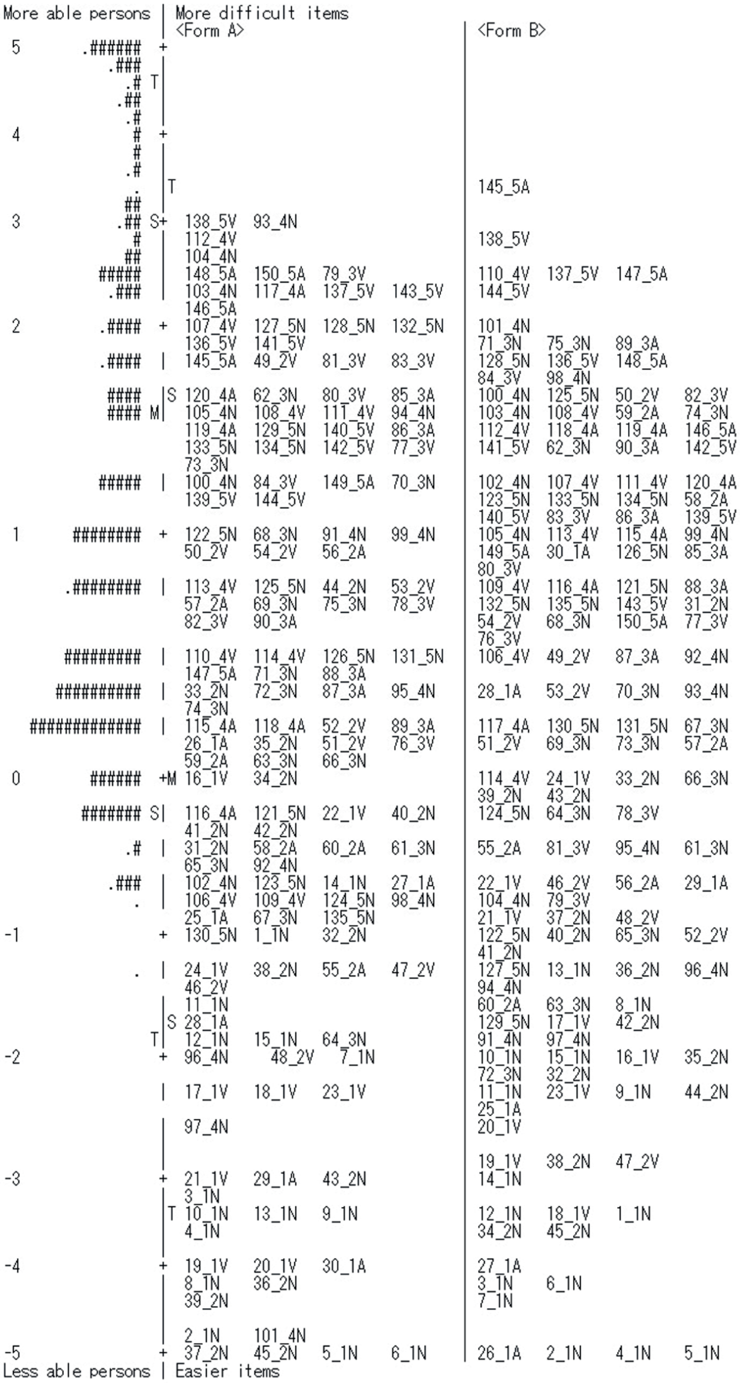
In order to statistically examine the homogeneity of variance of item difficulty between the two forms, Levene's test was performed. The results showed that the null hypothesis of equal variances was not rejected ($F = 0.160$, $p = .689$), indicating that the spread of item difficulties may be acceptably equal between the two forms. Table 3 shows the results of the subsequent *t*-tests (2-tailed) which examined the mean Rasch item difficulties[4] of each level. No statistically significant differences were detected for any level. Cohen's effect size (*d*) was below .20 for every level, indicating small differences between the two forms (Cohen, 1988, 1992). This may indicate that the two forms are statistically equivalent.

[4] Rasch item difficulty and person ability are expressed in logit (log odds unit). Logit is the unit of measurement on an interval scale to which raw scores are transformed by the Rasch model. In this study, the value of 0 logits is allocated to the mean item difficulty. Larger numbers indicate more difficult items and more able persons, and *vice versa*.

Table 3. Comparison of the item difficulty between the two equivalent forms

	Form A		Form B		*t*	*d.f.*	*p*	*d*
Level	*M*	*S.D.*	*M*	*S.D.*				
1000	−2.44	1.57	−2.56	1.79	0.28	58	.783	0.07
2000	−0.75	1.86	−0.71	1.37	0.10	58	.917	0.02
3000	0.71	0.83	0.55	1.03	0.68	58	.421	0.17
4000	0.53	1.70	0.67	1.03	0.38	58	.706	0.10
5000	1.22	1.06	1.11	1.12	0.40	58	.689	0.10
Total	**−0.14**	**1.95**	**−0.19**	**1.85**	**0.20**	**298**	**.844**	**0.03**

Another way of investigating the degree of equivalence of the two forms is to examine the item difficulty hierarchy for each form. This may be addressed by looking at a Rasch person-item map (or often called a Wright map), which displays both persons in terms of ability and items in terms of difficulty on a Rasch interval scale. Figure 5 is a person-item map for the two new forms. The far left of this figure shows a Rasch logit scale with the mean item difficulty being 0. This figure has two distributions on the logit scale: persons on the left and items on the right. More able persons and more difficult items are located towards the top and less able persons and less difficult items are located towards the bottom. For the person distribution, each "#" represents two persons and each "." represents one person. For the item distribution, the items of Form A are shown in the left and those of Form B on the right. Each number indicates the unique item number and the subsequent number and letter indicate the word level and the part of speech, respectively. For example, 138_5V means that the unique item number is 138, its word level is 5000, and its part of speech is verb. The two distributions (person and item) are interrelated in that a person has a 50% probability of succeeding on an item located at the same point on the logit scale. This person's success probability increases for items located lower than that point, and *vice versa*. Figure 5 shows that there are few gaps in the item difficulty hierarchy and the item difficulties are largely evenly distributed between Forms A and B.



```
More able persons | More difficult items
                  |  <Form A>                                | <Form B>
5        .######  +                                          |
            .###  |                                          |
              .# T|                                          |
             .##  |                                          |
              .#  |                                          |
4              #  +                                          |
               #  |                                          |
              .#  |                                          |
               .  |T                                         | 145_5A
              ##  |                                          |
3            .## S+  138_5V    93_4N                         |
               #  |  112_4V                                  | 138_5V
              ##  |  104_4N                                  |
           #####  |  148_5A    150_5A    79_3V               | 110_4V    137_5V    147_5A
            .###  |  103_4N    117_4A    137_5V    143_5V    | 144_5V
                     146_5A                                  |
2          .####  +  107_4V    127_5N    128_5N    132_5N    | 101_4N
                     136_5V    141_5V                        | 71_3N     75_3N     89_3A
           .####  |  145_5A    49_2V     81_3V     83_3V     | 128_5N    136_5V    148_5A
                                                             | 84_3V     98_4N
            ####  |S 120_4A    62_3N     80_3V     85_3A     | 100_4N    125_5N    50_2V     82_3V
            #### M|  105_4N    108_4V    111_4V    94_4N     | 103_4N    108_4V    59_2A     74_3N
                     119_4A    129_5N    140_5V    86_3A     | 112_4V    118_4A    119_4A    146_5A
                     133_5N    134_5N    142_5V    77_3V     | 141_5V    62_3N     90_3A     142_5V
                     73_3N                                   |
           #####  |  100_4N    84_3V     149_5A    70_3N     | 102_4N    107_4V    111_4V    120_4A
                     139_5V    144_5V                        | 123_5N    133_5N    134_5N    58_2A
                                                             | 140_5V    83_3V     86_3A     139_5V
1       ########  +  122_5N    68_3N     91_4N     99_4N     | 105_4N    113_4V    115_4A    99_4N
                     50_2V     54_2V     56_2A               | 149_5A    30_1A     126_5N    85_3A
                                                             | 80_3V
       .########  |  113_4V    125_5N    44_2N     53_2V     | 109_4V    116_4A    121_5N    88_3A
                     57_2A     69_3N     75_3N     78_3V     | 132_5N    135_5N    143_5V    31_2N
                     82_3V     90_3A                         | 54_2V     68_3N     150_5A    77_3V
                                                             | 76_3V
       #########  |  110_4V    114_4V    126_5N    131_5N    | 106_4V    49_2V     87_3A     92_4N
                     147_5A    71_3N     88_3A               |
      ##########  |  33_2N     72_3N     87_3A     95_4N     | 28_1A     53_2V     70_3N     93_4N
                     74_3N                                   |
   #############  |  115_4A    118_4A    52_2V     89_3A     | 117_4A    130_5N    131_5N    67_3N
                     26_1A     35_2N     51_2V     76_3V     | 51_2V     69_3N     73_3N     57_2A
                     59_2A     63_3N     66_3N               |
0         ######  +M 16_1V     34_2N                         | 114_4V    24_1V     33_2N     66_3N
                                                             | 39_2N     43_2N
         ####### S|  116_4A    121_5N    22_1V     40_2N     | 124_5N    64_3N     78_3V
                     41_2N     42_2N                         |
              .#  |  31_2N     58_2A     60_2A     61_3N     | 55_2A     81_3V     95_4N     61_3N
                     65_3N     92_4N                         |
            .###  |  102_4N    123_5N    14_1N     27_1A     | 22_1V     46_2V     56_2A     29_1A
               .  |  106_4V    109_4V    124_5N    98_4N     | 104_4N    79_3V
                     25_1A     67_3N     135_5N              | 21_1V     37_2N     48_2V
-1                +  130_5N    1_1N      32_2N               | 122_5N    40_2N     65_3N     52_2V
                                                             | 41_2N
               .  |  24_1V     38_2N     55_2A     47_2V     | 127_5N    13_1N     36_2N     96_4N
                     46_2V                                   | 94_4N
                  |  11_1N                                   | 60_2A     63_3N     8_1N
                  |S 28_1A                                   | 129_5N    17_1V     42_2N
                 T|  12_1N     15_1N     64_3N               | 91_4N     97_4N
-2                +  96_4N       48_2V    7_1N               | 10_1N     15_1N     16_1V     35_2N
                                                             | 72_3N     32_2N
                  |  17_1V     18_1V     23_1V               | 11_1N     23_1V     9_1N      44_2N
                                                             | 25_1A
                  |  97_4N                                   | 20_1V
                  |                                          |
                  |                                          | 19_1V     38_2N     47_2V
-3                +  21_1V     29_1A     43_2N               | 14_1N
                  |  3_1N                                    |
                  |T 10_1N     13_1N     9_1N                | 12_1N     18_1V     1_1N
                  |  4_1N                                    | 34_2N     45_2N
                  |                                          |
-4                +  19_1V     20_1V     30_1A               | 27_1A
                  |  8_1N      36_2N                         | 3_1N      6_1N
                  |  39_2N                                   | 7_1N
                  |                                          |
                  |  2_1N      101_4N                        |
-5                +  37_2N     45_2N     5_1N      6_1N      | 26_1A     2_1N      4_1N      5_1N
Less able persons | Easier items
```

Figure 5. Rasch person-item map of the two new forms

Notes: M = Mean, S = 1SD, T = 2SD, N = Noun, V = Verb, A = Adjective.

Validation

This section discusses the validity of the two new forms of the VLT from the five aspects of construct validity: content, substantive, structural, generalizability, and external. Construct validity is a unified concept that may be examined through the provision of evidence from various distinct aspects (e.g., Messick, 1989). In this study, the new forms of the VLT were validated based on Messick's (1989, 1995) framework because it has been accepted as a useful means of validation by researchers in language testing (Bachman, 1990, 2000; Bachman & Palmer, 1996; Chapelle, 1999; McNamara, 2006; Read & Chapelle, 2001) as well as in psychology and education (e.g., APA, AERA, & NCME, 1999).

Content aspect of construct validity

The content aspect of construct validity aims to clarify "the boundaries of the construct domain to be assessed" (Messick, 1995, p. 745). This aspect addresses content relevance, representativeness, and technical quality of the items (Smith, Jr. 2004). Content relevance refers to the relationship between the test items and the construct being measured (receptive knowledge of the form-meaning relationships of words). The new forms of the VLT were considered to be representative of the construct domain, because (1) the target words were selected based on a stratified random sampling method from each 1000-word frequency band, and (2) the ratio of the three parts of speech reflected actual language use (Noun : Verb : Adjective = 3:2:1).

Representativeness may be empirically evaluated by examining the item strata which indicates the number of statistically different levels of item difficulty. It is derived using the following formula:

$$\text{Item strata} = (4\ G\mathit{item}+1)/3,$$

where G*item* is Rasch item separation. Item strata statistics need to be greater than 2.0 for useful tests, because "[i]f a sufficient (at least 2) number of item difficulty levels are unable to be identified, then one may have difficulty in interpreting the variable defined by the items" (Smith Jr., 2004, p. 106). Forms A and B showed the strata statistics of 6.85 and 7.12, respectively. This indicates that both forms have more than two statistically distinct difficulty levels, which can be taken as supportive evidence of their representativeness.

Another way of examining representativeness is to see whether there are gaps in the item difficulty hierarchy. Figure 5 shows that there are few gaps in the item difficulty hierarchy between 3 and –5 logits, indicating a high degree of representativeness in terms of item difficulty.

Technical quality may be investigated by examining the degree to which the empirical data fit the Rasch model (Smith Jr., 2004). The technical quality of the two new forms of the VLT should be high because the items were selected from the ones that fitted the Rasch model (see the *Development of two equivalent forms* section).

Technical quality was empirically examined by inspecting item correlations and fit statistics. First, the point-measure correlation[5] (correlation between the observations on an item and the corresponding person ability estimates) was examined to see whether the items are aligned in the same direction as the latent variable. The point-measure correlation measures the degree to which more able persons scored higher (or less difficult items were scored higher). The values range between −1 and 1, and the items with negative values need to be inspected. The results showed that all items showed positive point-measure correlations.

Second, Rasch outfit and infit *t* statistics were inspected for fit analysis. Misfit items are the ones with *t* values greater than 2 (underfit) or smaller than −2 (overfit). Underfit is usually taken as a more serious problem than overfit because it indicates that the quality of the items is degraded by many unexpected responses that do not conform to the Rasch model. The analysis revealed that 15 items showed infit or outfit *t* values larger than 2. This represents only a 5% misfit rate (15 out of 300 items), which may be expected given the nature of Type I (alpha .05) error rate.

Finally, a qualitative inspection was made of the 15 misfit items. The most misfitting item was 84_3V on Form A (infit $t = 4.4$, outfit $t = 5.6$). Table 4 shows the statistics of the choices (the target definition is "try to win" and the correct answer is "compete"). This table shows that the average abilities of those who chose "bargain" (1.30) and "dedicate" (1.77) approached the average ability of those who chose the correct answer "compete" (1.82). However, "bargain" and "dedicate" seem to be semantically different from "try to win." In addition, the correct answer (compete) was chosen by the largest number of people with the highest average ability. This may indicate that this item is unlikely to cause a serious problem. The other 14 misfit items were inspected in this way, and no serious problem was found with any items.

Table 4. Choice statistics for 84_3V on Form A

Option	compete	assault	bargain	dedicate	nominate	restrain
% chosen	53.2	2.4	11.1	7.9	22.2	3.2
Ave. ability (logits)	1.82	0.99	1.30	1.77	1.05	0.47

[5] The point-measure correlation, rather than point-biserial correlation, was used because the former is more robust with missing data than the latter (Linacre, 2016b, p. 536).

Another issue relating to technical quality is local independence: the Rasch model requires that all items be independent of each other (e.g., Bond & Fox, 2015). The VLT may violate local independence because three items share the same six choices in each cluster. One way of investigating local independence is fit statistics. It is generally indicated by overfit (outfit $t < -2$ or infit $t < -2$). No items had outfit t of smaller than −2. In terms of infit statistics, 14 items (4.7%) showed overfit values, but no two items were included in the same cluster. Another way of investigating local independence was analyzing standardized residual correlations (correlations of the residuals which are not explained by the Rasch model). Linacre (2016b, p. 399) suggests that a correlation of around 0.7 and above signals dependency. The results showed that no item pairs had residual correlations of 0.7 or above. Three item pairs had a residual correlation of larger than 0.6 (0.69, 0.67, and 0.61) which means more than 36% of their variance were in common, but none of them were presented in the same cluster. This indicates that the new forms of the VLT items may be acceptable in terms of local independence.

Substantive aspect of construct validity

The substantive aspect of construct validity refers to "theoretical rationales for the observed consistencies in test responses […] along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks" (Messick, 1995, p. 745). This aspect may be evaluated by examining whether the empirical item hierarchy is presented as predicted by theoretical argument and whether each person's response pattern is consistent with that item hierarchy (Smith Jr., 2004).

For the item hierarchy, it was hypothesized that higher-frequency-level items would be easier than lower-frequency-level items, although this tendency is less clear for lower-frequency levels (see Beglar, 2010; McLean, Kramer, & Beglar, 2015 for support of this hypothesis). Figures 6 and 7 illustrate the mean item difficulty estimates and their 95% confidence intervals for the five word frequency levels in Forms A and B, respectively. These figures indicate a general tendency that items become more difficult as word frequency decreases, although this tendency is less transparent at the 3000-, 4000-, and 5000-word levels.

A one-way ANOVA was performed to examine whether the mean item difficulties were statistically different between the word frequency levels. The results showed a statistically significant difference for both Forms A ($F(4, 145) = 30.6$, $p = .000$) and B ($F(4, 145) = 39.3, p = .000$). Table 5 presents the results of a Tukey's post-hoc test, indicating that the 1000- and 2000-word levels were statistically different from the other levels, but no statistically significant difference was found between 3000-, 4000-, and 5000-word levels. One reason why there may be little difference between knowledge at the 3000 to 5000 frequency levels is that a lack

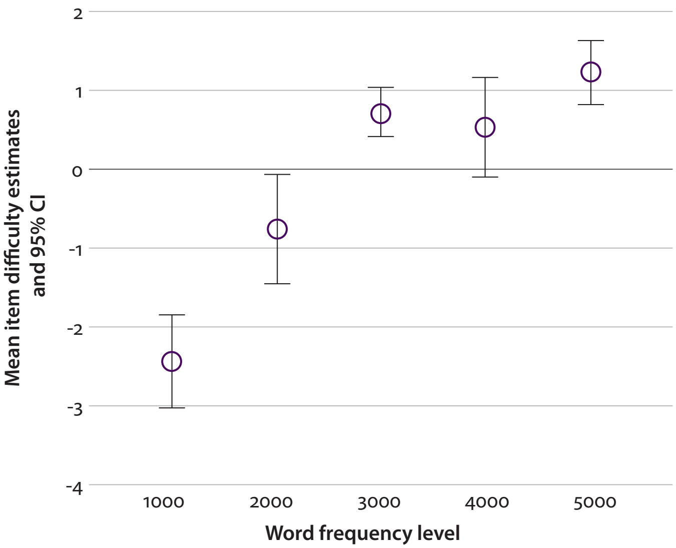


Figure 6. Mean item difficulties for Form A

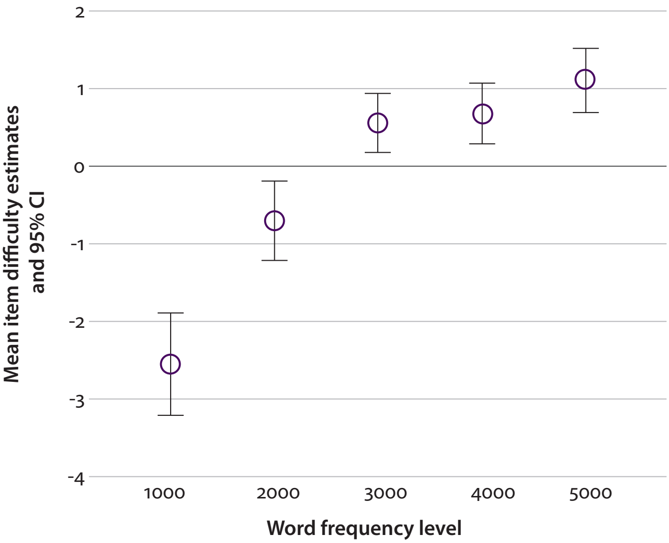


Figure 7. Mean item difficulties for Form B

of L2 input in the EFL context may reduce the effects of lexical frequency for less frequent words. For example, there may be sufficient lexical input within the classroom and course books to differentiate knowledge of the highest frequency words (e.g., 1000 level words such as boy, pull, beautiful are better known than 2000 level words such as career, operate, advanced). However, the same may not always hold true of slightly less frequent words (e.g., 4000 level words such as auction, roast, and delicate may be known to a similar degree as 5000 level words such as foam, abolish, and extinct), because words at the 4000 level may not always be encountered much more often than those at the 5000 word level in the EFL context.

Table 5. A *p*-value matrix for pairwise comparisons in mean item difficulties

Form A					**Form B**			
Level	**1000**	**2000**	**3000**	**4000**	**1000**	**2000**	**3000**	**4000**
2000	.000				.000			
3000	.000	.001			.000	.002		
4000	.000	.007	.987		.000	.001	.997	
5000	.000	.000	.668	.357	.000	.000	.465	.683

Another way to investigate the substantive aspect of construct validity is to examine the consistency of each person's response pattern with the item hierarchy. More specifically, Rasch person fit statistics were calculated for the two forms of the VLT. Person fit examines the degree of match between the observed responses and the theoretical model that requires a person of a given ability to have a greater probability of a higher rating on easier items than on more difficult items (Smith Jr., 2004). As with item fit, a misfit person was defined as outfit $t > 2.0$ or infit $t > 2.0$ (underfit), or outfit $t < -2.0$ or infit $t < -2.0$ (overfit). The results showed the misfit rate of less than 5% both for Forms A (4.1%) and B (4.7%), which is expected to occur by chance. This indicates that the test-takers' response pattern corresponded to the modelled difficulty order. This may be taken as supportive evidence for the substantive aspect of construct validity.

Structural aspect of construct validity

The structural aspect of construct validity "appraises the fidelity of the scoring structure to the structure of the construct domain at issue" (Messick, 1995, p. 745). The evaluation of this aspect may be addressed by examining the unidimensionality (the degree to which a test measures one attribute at a time) of the intended structure, because a unidimensional measure allows a straightforward scoring method (Smith Jr., 2004; Wolfe & Smith Jr., 2007). Linacre (1995) suggested that dimensionality may be addressed by (1) item correlations, (2) fit statistics, and (3) principal components analysis (PCA) of standardized residuals without rotation. In terms of item correlations and fit statistics, the new forms of the VLT may be acceptably unidimensional (see Section 3.3.2 *Content aspect of construct validity* for a detailed discussion).

The PCA of standardized residuals was performed in order to examine whether there was only a small amount of variance in the residuals accounted for by other components (contrasts) than the Rasch model which extracts the first major component in the observations. Figure 8 presents the scree plot for the VLT. This figure shows that (1) the first dimension (Rasch model) accounted for 39.8% of the variance in the residuals, (2) the factor sensitive ratio (Wright & Stone, 2004) (eigenvalue of the 1st contrast divided by that of the Rasch model) is only 3.9%, and (3) the eigenvalues of other contrasts seem to reach an asymptote at the first contrast (see Stevens, 2002; Wolfe & Smith Jr., 2007 for a detailed discussion). This may be taken as positive evidence for unidimensionality of the new forms of the VLT. It should be noted here that the eigenvalue of the first contrast is above the chance level of around 2 (Linacre & Tennant, 2009; Raîche, 2005). Tables 6 and 7 show the items with substantial positive and negative loadings of greater than .40 and smaller than –.40 on the first contrast. These items do not seem to have a common meaning to constitute a different dimension. Taken together, the new forms of the VLT are likely to measure the unidimensional construct, that is, receptive knowledge of the form-meaning connections of words.

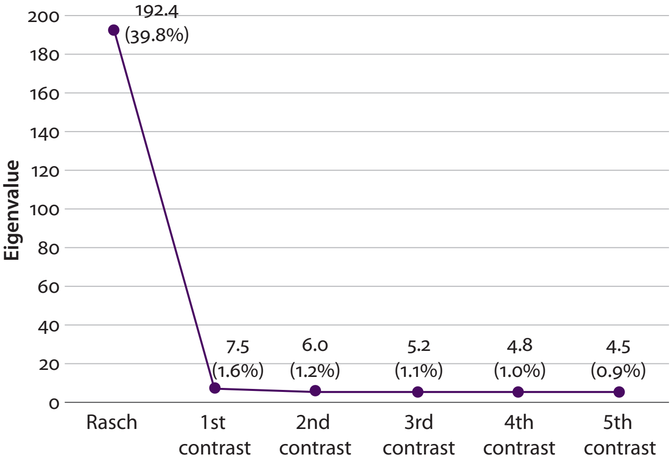


Figure 8. Scree plot for the VLT

Table 6. Items with positive loadings (>.40) on the first contrast of residuals

Form	Item	Loading	Difficulty	Target word	Definition
B	82_3V	.58	1.60	Persist	Continue to happen
B	140_5V	.48	1.26	Intrude	Enter without permission
A	86_3A	.45	1.34	Mortal	Can die
B	139_5V	.44	1.26	Notify	Announce
B	121_5N	.42	0.80	Mustache	Hair on your upper lip

Table 7. Items with negative loadings (<−.40) on the first contrast of residuals

Form	Item	Loading	Difficulty	Target word	Definition
B	43_2N	-.50	0.01	Envelope	Cover for letters
B	36_2N	-.45	-1.12	Average	Middle number
B	44_2N	-.43	-2.27	Cap	Kind of hat
B	74_3N	-.42	1.35	Heritage	History

Generalizability aspect of construct validity

The generalizability aspect of construct validity deals with "the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks" (Messick, 1995, p. 745). This aspect may be approached by examining the extent to which item difficulty and person ability estimates are invariant within the measurement error across measurement contexts such as different groups of examinees, time, or tasks (Andrich, 1988; Smith Jr., 2004; Wolfe & Smith Jr., 2007; Wright & Stone, 1979). Wolfe and Smith Jr. (2007) divided this aspect into four subcategories: item calibration invariance (stability of item difficulty estimates), person measure invariance (stability of person ability estimates), reliability (stability of measures across instrument and scoring designs), and invariance across administrative contexts.

(1) *Item calibration invariance.* The invariance of item calibrations refers to "the degree to which item calibrations maintain their meaning and interpretability […] across groups of respondents and across time" (Wolfe & Smith Jr., 2007, p. 215). This was investigated by analyzing uniform differential item functioning (DIF), an indication of unexpected behavior by items showing that item calibrations vary across samples by more than the modelled error.

First, the DIF analysis was performed in order to examine whether the item calibrations from male ($N = 51$) and female ($N = 196$) test-takers varied widely for each of the three sections.[6] A Mantel-Haenszel test (Mantel & Haenszel, 1959) revealed that no statistically significant DIF was detected for any items ($\alpha = .05$).

DIF was also investigated in terms of test-takers' native language; that is, whether the item difficulty estimates from Japanese ($N = 148$), Spanish ($N = 62$), and Chinese ($N = 40$) learners varied widely. Through a Mantel-Haenszel approach, significant DIF ($\alpha = .05$) was found for 10 items in total (Tables 8 and 9).[7] Table 8 shows that the

[6] Three test-takers without a response to gender were deleted from the analysis.

[7] The L1 influence might be underestimated, because a simulation study indicates that DIF analyses require more than 200 respondents per group for obtaining adequate (>80% power) performance (Scott et al., 2009).

four words (*mortal, intrude, persist, altitude*) are more difficult for Japanese learners than Spanish learners, and *vice versa* for the other four words (*pit, crown, random, envelope*). This may have been because L1 knowledge of cognates and loan words gave advantages to one group of learners over another for several words. However, the analysis indicated that it is unlikely that the VLT favors one particular L1 group over another. In addition, significant DIF was found only for 10 out of 900 (1.1%) instances (300 items by three L1 combinations). This may be taken as positive evidence of the VLT's generalizability.

Table 8. DIF analysis for Japanese and Spanish learners

Form	Item	Target word	Difficulty estimates for Japanese	Difficulty estimates for Spanish	*Chi-square*	*p*
A	86_3A	mortal	1.86	−2.09	7.53	.006
B	62_3N	pit	1.16	2.77	7.10	.007
A	31_2N	crown	−0.81	1.35	5.95	.015
B	140_5V	intrude	1.78	−1.57	4.83	.028
A	87_3A	random	−0.05	1.92	4.24	.040
B	43_2N	envelope	−0.42	1.72	4.17	.041
B	82_3V	persist	2.61	−1.57	4.17	.041
B	133_5N	altitude	1.78	−0.28	4.17	.041

Table 9. DIF analysis for Japanese and Chinese learners

Form	Item	Target word	Difficulty estimates for Japanese	Difficulty estimates for Chinese	*Chi-square*	*p*
B	115_4A	credible	1.40	−0.53	4.81	.028
A	116_4A	amateur	−0.67	2.15	4.23	.040

(2) *Person measure invariance.* The invariance of person ability estimates was examined by analyzing differential person functioning (DPF), an indication of unexpected behavior by persons. Specifically, it was examined whether person ability estimates from different parts of speech (noun, verb, and adjective) fell within a measurement error. The DPF analysis was performed through a *t*-test approach with reference to the baseline measures (ability estimates from all responses) and significant DPF ($\alpha = .05$) was detected for 21 (12 Japanese, 4 Spanish, and 5 Chinese persons) out of 750 cases (250 test-takers by 3 parts of speech). Out of the 21 cases, significant DPF was found for 11, 4, and 6 persons from the estimates of noun, verb, and adjective items, respectively. For every part of speech, the DPF rate (DPF persons divided by 250 test-takers) is below the chance level of 5%. This

indicates that different part-of-speech items contribute to estimation of the unidimensional construct.

(3) *Reliability.* A third way of investigating the generalizability aspect of construct validity is to examine the degree of reliability. Rasch analysis provides two types of reliability: person and item reliability. Person reliability is equivalent to traditional reliability coefficients such as Cronbach's alpha, KR-20, and the Generalizability coefficient. Rasch item reliability, which has no traditional equivalent, addresses the degree to which item difficulties are reproducible. Rasch analysis also presents person and item separation estimates which are linear and range from zero to infinite. The conventional reliability estimates are non-linear and suffer from ceiling effects within the range between zero and one (Smith Jr., 2004). Table 10 presents the Rasch reliability and separation estimates for Forms A and B. The results showed that the reliability estimates were .96 and separation estimates were 4.72 and above. This indicates that the person ability and the item difficulty estimates are highly reproducible.

Table 10. Rasch reliability and separation estimates

	No. of items	**No. of test-takers**	**Person reliability**	**Person separation**	**Item reliability**	**Item separation**
Form A	150	127	.96	4.89	.96	4.72
Form B	150	123	.96	5.09	.96	4.81

(4) *Invariance across administrative contexts.* A final way of evaluating the generalizability aspect is to examine the stability of performance across administrative contexts. For future use of the VLT, person ability will be estimated based on the performance on the 150 items in either Form A or B, without using a common-item linking method and intentional missing data as designed for the present research. Thus, administrative invariance was evaluated by examining the degree to which the person ability estimates from the short version (150 items) were consistent with those from the long version (168 items based on the common-item linking design). A paired *t*-test was performed for each section in order to investigate whether a statistically significant difference was found between the person ability estimates from these two versions. Table 11 presents the mean Rasch person ability estimates in logits for the two versions, *t*-statistics, and Cohen's *d*. This table shows that no significant difference was found between the short and long versions for both forms, and the effect size was negligible.

Table 11. Rasch person measures, *t*-statistics, and effect size between the short and long versions for the three sections

Form	No. of test-takers	Short version		Long version		*t*	*d.f.*	*p*	*d*
		M	*S.D.*	*M*	*S.D.*				
A	127	1.51	1.70	1.57	1.64	0.28	126	.780	0.04
B	123	1.54	1.83	1.50	1.78	0.17	122	.867	0.02

External aspects of construct validity

The external aspect refers to "the extent to which the test's relationships with other tests and nontest behaviors reflect the expected high, low, and interactive relations implied in the theory of the construct being assessed" (Messick, 1989, p. 45). In order to examine the relationship with another test measuring the related construct, a passive recall format (writing a meaning to the target word) was created. The passive recall format measured knowledge of 30 words in 10 clusters (2 clusters from each level) which consisted of 5 noun, 3 verb and 2 adjective clusters. It was hypothesised that the scores from the passive recall format and those from the VLT would be moderately correlated (Pearson's r of around .6) based on Laufer and Goldstein (2004) who found the correlation coefficients of $r = .58$ (passive recall and passive recognition) and $r = .65$ (passive recall and active recognition). In order to test this hypothesis, the passive recall format and the VLT format with the same 10 clusters were administered to 31 (27 male, 4 female) Japanese university students in a paper-based format. The test takers were first-year engineering students and their proficiency level as measured by TOEIC was $M = 361.5$, $S.D. = 26.2$. For the passive recall format, the students gave Japanese translations of the target words, and the translations were scored as correct (1 point) or incorrect (0 point) by two Japanese native speaking university teachers of English. The two raters had a discussion about the 13 responses with inconsistent scores to reach agreement. The results showed that the correlation coefficient (Pearson's r) between the two versions was .649, which may be taken as supportive evidence for the hypothesis.

Table 12 presents descriptive statistics of the scores from the two formats. This table shows that the VLT format yielded 1.61 times higher scores than the passive recall format. This is in line with Laufer and Goldstein's (2004) finding that the scores from the recognition formats were 1.42 to 1.83 times higher than those from the passive recall format. A close look at the response pattern indicates that the higher scores in the VLT format do not seem to be due to random guesswork: some words had many more correct responses in the VLT format, while others did not. For example, the word *lone* was answered correctly by 5 students in the recall format, but 22 got it correct in the VLT format. A subsequent interview revealed that the presentation of the choices (definitions) helped students to recognize familiar words such as *lonely* and *alone*. The words *glance* and *spectator* were other words that had 15 more correct responses in the VLT format. These higher scores indicate that the VLT may be better able to tap into partial vocabulary knowledge than the recall format.

Table 12. Descriptive statistics for the passive recall and the VLT formats

	M	SD	Max	Min
VLT	16.9	3.7	22	8
Recall format (30)	10.5	3.0	16	6

Discussion

This study described the development and validation of two new forms of the VLT. It provided empirical evidence for the equivalence of the two new forms of the VLT and initial validity evidence for them. The new forms improve on the earlier versions in three ways. First, the inclusion of 1000 word levels allows teachers and researchers to measure knowledge of the vocabulary that likely has the greatest impact on a learner's ability to communicate in English. Earlier studies examining the lexical coverage of different types of discourse have shown that knowing a greater proportion of words in spoken (Stæhr, 2009; van Zeeland & Schmitt, 2013) and written discourse (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Schmitt, Jiang, & Grabe, 2011) increases the potential that language will be understood. The most frequent 1000 word families together with proper nouns and interjections accounts for 86.52% of movies (Webb & Rodgers, 2009a), 85.11% of television programs (Webb & Rodgers, 2009b), 83.25% of text written for children (Webb & Macalister, 2013), 91.06 of graded readers (Webb & Macalister, 2013), 87.54% of academic spoken English (Dang & Webb, 2014), and from 64.74–88.00% of English proficiency test passages (Webb & Paribakht, 2015). Thus, because the most frequent 1000 word families account for by far the largest proportion of English vocabulary, measuring this word frequency level on its own has great value.

Second, the items from the new forms of the VLT were sourced from Nation's (2012) BNC/COCA word frequency lists. These lists were derived from mega-corpora that were designed to reflect current English in the United Kingdom and the United States. Thus, the frequency of the items in the test should provide a reasonable representation of the likelihood that vocabulary will be encountered in these contexts, as well as within a large proportion of language learning materials.

The new test forms should therefore provide face validity for test takers over the next few years.

Third, the addition of 4000 word levels allows the test to better reveal gaps in lexical knowledge than in the earlier versions of the test. Measuring knowledge of five sequenced word frequency levels should help teachers (and learners) to see the extent of vocabulary learning progress. Moreover, the new forms of the VLT should more clearly reveal when there is a lack of learning progress from year to year that should in turn help to evaluate the efficacy of the vocabulary learning programs within institutions. Because many learners in EFL contexts struggle to make progress with their lexical development, evaluating the efficacy of vocabulary learning programs is of particular importance (Webb & Chang, 2012).

Fourth, both forms of the VLT developed in this study are freely available. Form A is included in the appendix, and Form B of the VLT is freely available in paper-based and electronic formats at Stuart Webb's homepage <http://www.edu.uwo.ca/faculty-profiles/stuart-webb.html>

The creation of two equivalent forms should help users to avoid potential test retake effects, where test takers repeated use of the same test may lead to an increase in scores. In addition, the electronic form of the test has two useful features. First, it provides test takers with some feedback about their performance on the test. This includes a brief explanation of their score and the level that they should focus their vocabulary learning. Second, if teachers wish to have their students take the test in a computer lab or at home, they can provide them with an email address that students can enter at the end of the test, and test results will be sent to that address. In that way, teachers can quickly collect their students' test scores.

Finally, it is important to note that while there are available tests that include measures of the 1000 word level such as Nation and Beglar's (2007) and Coxhead, Nation, and Sim's (2015) Vocabulary Size Test, these tests do not have a sufficient number of items to reliably measure knowledge of individual word levels, as this is not their purpose. Vocabulary size tests are designed to provide a valid and reliable measure of lexical knowledge as a whole rather than individual frequency levels.

Interpreting scores

As mentioned earlier, when interpreting scores, it is the scores for the individual levels of the VLT that are meaningful rather than the scores for all levels combined. Because higher frequency words have greater value than lower frequency words, when interpreting scores, users should look at scores on the levels according to their frequencies. The highest frequency level that has not been mastered, should be where attention is focused for further learning. In the original form of the VLT, Nation (1983) recommended that scores less than 66% for a level indicated that words from that level needed further study. In their updated forms of the VLT, Schmitt, Schmitt, & Clapham (2001) suggested a higher threshold for mastery of a level. They recommended that if test takers scores were 26/30 (87%) or higher, they had achieved mastery of that level and might then focus on learning words from the next level. However, subsequently a lower cutting point of 24/30 (80%) was suggested by Schmitt as being sufficient for mastery of a level (Xing & Fulcher, 2007). Xing and Fulcher (2007) note that in all discussions of score interpretation of the VLT, the mastery cutting point appears to have been arbitrary.

We propose that the score for mastery of each level should depend to some degree on the level; at the 1000, 2000, and 3000 levels we recommend a cutting point of 29/30, while at the 4000 and 5000 levels the cutting point might remain at 24/30. The reason for the higher cutting point for the first three levels is that these words account for such a large percentage of English, they provide the foundation for further lexical and language development. For example, knowing the most frequent 3000 word families accounts for 95% of many types of spoken discourse, and this percentage of lexical coverage may be sufficient for comprehension of spoken input (van Zeeland & Schmitt, 2013). Working towards achieving mastery of the most frequent 3000 word families thus has great value. When considering the individual levels, the most frequent 1000 word families make up by far the greatest proportion of English. There is therefore greater value in helping to achieve near perfect knowledge of their form-meaning connections before moving on to the 2000 level. The same argument holds true at the 2000 and 3000 levels; although the proportion of language that these levels represent is far smaller than the 1000 level, they still represent a relatively large percentage of spoken and written text, and so a very high cutting point makes sense.

It is also important to note that the VLT measures relatively shallow knowledge of a word and that there is much more to knowing words than simply recognizing their form-meaning connections. It may thus be better to be cautious and use a higher cutting point for mastery of the highest frequency levels. Because mid frequency vocabulary begins from the 4000 word level (Schmitt & Schmitt, 2014), a lower cutting point such as 24/30 might still be considered appropriate.

Limitations

The new forms of the VLT include some of the limitations of the earlier versions. First, as with all tests, validation should be considered an ongoing process. Although the development and validation of the new forms of the VLT went beyond that of several earlier vocabulary tests, there is still value in further examining its validity with learners in different contexts. Second, it is important to note what the VLT can and cannot measure. It can measure the extent to which test takers are able to recognize the form-meaning connections of words at five different word frequency levels. Thus, it is a measure of receptive vocabulary knowledge indicating the degree to which test takers may be able to understand the meanings of words that they encounter in written text. However, because it does not measure productive vocabulary knowledge, it does not measure the extent to which test takers can produce L2 words. Similarly, it does not indicate the degree to which test takers can use words at different frequency levels, nor does it indicate the degree to which test takers have knowledge of other aspects of vocabulary knowledge such as collocation, word parts, and polysemy. There is still a need for the further development and validation of tests that isolate and measure these aspects of vocabulary knowledge (Nation & Webb, 2011; Webb, 2013; Webb & Sasao, 2013). Finally, as mentioned earlier, the VLT should not be considered a test of vocabulary size because most learners even at the beginner level are likely to know some words that are lower in frequency than the 5000 word level.

Conclusion

This study presents initial evidence supporting the validity of two new equivalent forms of the VLT. The updated VLT improves on earlier versions by measuring knowledge of the most important words in English: the first five 1000-word frequency levels from Nation's (2012) BNC/COCA word lists. Although we believe that the updated VLT will be of value to teachers, learners, and researchers, there is always a need for the development and validation of new tests of lexical knowledge. In particular, there would be value in developing tests designed to measure different aspects of knowledge apart from form-meaning connection such as collocation and polysemy. It would also be useful to develop a test designed to measure smaller frequency bands within the most frequent 1000 word families or the 800 lemmas that make up the Essential Word List.

With the development of new tests, it is also important to take the time to provide validity evidence in support of their use. Although this can be a relatively long process (development and validation of the new forms of the VLT in this study took about four years), test users should then have the confidence that the test is accurately measuring what it was intended to measure. We urge users to take the time to determine whether there is sufficient evidence to support the use of the vocabulary tests that they use.

References

Andrich, D. (1988). *Rasch models for measurement*. Beverly Hills, CA: Sage. doi: 10.4135/9781412985598

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, 17(1), 1–42.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

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

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York, NY: Routledge.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, 19, 254–272. doi: 10.1017/S0267190599190135

Cohen, J. (1988). *Statistical power analysis for the behavioral science* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159. doi: 10.1037/0033-2909.112.1.155

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213–238.

Coxhead, A., Nation, I. S. P., & Sim, D. (2015). Measuring the vocabulary size of native speakers of English in New Zealand secondary schools. *New Zealand Journal of Educational Studies*, 50(1), 121–135. doi: 10.1007/s40841-015-0002-3

Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes*, 33(1), 66–76. doi: 10.1016/j.esp.2013.08.001

Dang, T. N. Y., & Webb, S. (2016). Making an essential word list for beginners. In I. S. P. Nation, *Making and using word lists for language learning and testing* (pp. 153–167, 188–195). Amsterdam: John Benjamins. doi: 10.1075/z.208.15ch15

Hu, H. M., & Nation, P. (2000). What vocabulary size is needed to read unsimplified texts. *Reading in a Foreign Language*, 8, 689–696.

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement*, 1, 152–176.

Kremmel, B. (2016). Word families and frequency bands in vocabulary tests: Challenging conventions. *TESOL Quarterly*, 50(4), 976–987. doi: 10.1002/tesq.329

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, 54(3), 399–436. doi: 10.1111/j.0023-8333.2004.00260.x

Laufer, B., & Levitzky-Aviad, T. (2016). Computer Adaptive Test of Size and Strength. Retrieved from <http://catss.ga/>

Laufer, B., & Ravenhorst-Kalovski, G. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1), 15–30.

Linacre, J. M. (1995). Prioritizing misfit indicators. *Rasch Measurement Transactions*, 9(2), 422–423.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16(2), 878.

Linacre, J. M. (2003). Size vs. significance: Infit and outfit mean-square and standardized chi-square fit statistic. *Rasch Measurement Transactions*, 17(1), 918.

Linacre, J. M. (2016a). WINSTEPS® Rasch measurement computer program. Beaverton, Oregon: Winsteps.com.

Linacre, J. M. (2016b). *WINSTEPS® Rasch measurement computer programs User's Guide*. Beaverton, Oregon: Winsteps.com.

Linacre, J. M., & Tennant, A. (2009). More about critical eigenvalue sizes (variences) in standardized-residual principal components analysis (PCA). *Rasch Measurement Transactions*, 23(3), 1228.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719–748.

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

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly*, 3(1), 31–51. doi: 10.1207/s15434311laq0301_3

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests.

Meara, P., & Miralpeix, I. (2017). *Tools for researching vocabulary*. Bristol, UK: Multilingual Matters.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749. doi: 10.1037/0003-066X.50.9.741

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

Nation, I. S. P. (2012). The BNC/COCA word family lists. Retrieved from <http://www.victoria.ac.nz/lals/about/staff/paul-nation>

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

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

Raîche, G. (2005). Critical eigenvalue sizes in standardized residual principal components analysis. *Rasch Measurement Transactions*, 19(1), 1012.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.

Read, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Read, J., & Chapelle, C. (2001). A framework for second language vocabulary assessment. *Language Testing*, 18(1), 3–32. doi: 10.1191/026553201666879851

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, 95(1), 26–43. doi: 10.1111/j.1540-4781.2011.01146.x

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.

Schmitt, N., & Zimmerman, C. (2002). Derivative word forms: What do learners know? *TESOL Quarterly*, 36(2), 145–171.

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., … Sprangers, M. A. G. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of Clinical Epidemiology*, 62(3), 288–295. doi: 10.1016/j.jclinepi.2008.06.003

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology*, 8(33), 1–11.

Smith Jr., E. V. (2004). Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 93–122). Maple Grove, MN: JAM Press.

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, 31(04), 577–607. doi: 10.1017/S0272263109990039

Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

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. (2013). Depth of vocabulary knowledge. In C. Chappelle (Ed.), *Encyclopedia of Applied Linguistics* (pp. 1656–1663). Oxford, UK: Wiley-Blackwell.

Webb, S., & Sasao, Y. (2013). New directions in vocabulary testing. *RELC Journal*, 44(3), 263–278.

Webb, S. A. & Chang, A. C. -S., (2012). Second language vocabulary growth. *RELC Journal*, 43(1), 113–126. doi: 10.1177/0033688212439367

Webb, S., & Macalister, J. (2013). Is text written for children appropriate for L2 extensive reading? *TESOL Quarterly*, 47(2), 300–322. doi: 10.1002/tesq.70

Webb, S., & Nation, P. (2017). *How Vocabulary is Learned*. Oxford: Oxford University Press.

Webb, S., & Paribakht, T. S. (2015). What is the relationship between the lexical profile of test items and performance on a standardized English proficiency test? *English for Specific Purposes*, 38, 34–43. doi: 10.1016/j.esp.2014.11.001

Webb, S. & Rodgers, M. P. H. (2009a). The lexical coverage of movies. *Applied Linguistics*, 30(3), 407–427. doi: 10.1093/applin/amp010

Webb, S. & Rodgers, M. P. H. (2009b). The vocabulary demands of television programs. *Language Learning*, 59(2), 335–366. doi: 10.1111/j.1467-9922.2009.00509.x

Wolfe, E. W., & Smith Jr., E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part 2 – Validation activities. *Journal of Applied Measurement*, 8, 204–234.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Wright, B. D., & Stone, M. H. (2004). *Making measures*. Chicago, IL: Phaneron Press.

Xing, P., & Fulcher, G. (2007). Reliability assessment for two versions of Vocabulary Levels Tests. *System*, 35(2), 182–191.

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

Appendix 1. Form A of The Vocabulary Levels Test

This is test that looks at how well you know useful English words. Put a check under the word that goes with each meaning. Here is an example.

	game	island	mouth	movie	song	yard
land with water all around it						
part of your body used for eating and talking						
piece of music						

It should be answered in the following way.

	game	island	mouth	movie	song	yard
land with water all around it		✓				
part of your body used for eating and talking			✓			
piece of music					✓	

1,000 Word Level

	boy	rent	report	size	station	thing
how big or small something is						
place buses and trains go to						
young man						

	ear	gold	lake	letter	office	people
information sent to people						
men and women						
place for working						

	fellow	hat	ice	joke	light	system
funny story						
man or boy						
something worn on your head						

	date	forest	mistake	news	record	shop
latest information						
place with many trees						
something that is not right						

	bar	conversation	neighbor	rain	rubbish	shirt
person who lives nearby						
things that are thrown away						
type of clothing						

	continue	cook	phone	pull	sail	share
hold and move something toward yourself						
keep happening						
use together with others						

	enter	finish	happen	own	sing	worry
end						
go inside						
have something that is yours						

	arrive	collect	consider	glance	need	pack
look quickly at something						
reach the place you are going						
think about something						

	affordable	beautiful	boring	dry	rough	tall
higher than normal						
not flat						
not interesting						

	closed	dirty	empty	musical	orange	sad
having nothing						
not clean						
unhappy						

2,000 Word Level

	capital	**career**	**committee**	**exam**	**fence**	**option**
choice						
job						
test						

	guard	**lesson**	**library**	**license**	**monkey**	**soup**
food made with lots of water						
person who watches for danger						
place where many books are kept						

	brake	**crown**	**hero**	**language**	**mission**	**tale**
hat worn by a king or queen						
job						
things that stops a car						

	affair	**carrot**	**damage**	**desert**	**shelter**	**thief**
person who steals						
place that gives protection						
place with little rain						

	advice	**hobby**	**industry**	**soil**	**steak**	**storm**
bad weather						
earth						
things that you often enjoy doing						

	burst	**cheat**	**direct**	**operate**	**presume**	**wander**
believe something is true						
break open						
make something work						

	develop	**identify**	**improve**	**possess**	**provide**	**sew**
give						
have						
make better						

	complain	**increase**	**pray**	**produce**	**recognize**	**whip**
get larger						
know and remember						
make						

	curious	**defensive**	**energetic**	**nervous**	**various**	**wicked**
different kinds of things						
very bad						
wanting to know						

	advanced	**cruel**	**lone**	**stiff**	**typical**	**upset**
at a high level						
not kind						
single						

3,000 Word Level

	colleague	**fate**	**fee**	**hint**	**status**	**talent**
ability or skill						
clue						
person you work with						

	circuit	**clinic**	**format**	**origin**	**peak**	**routine**
place where you can see a doctor						
top						
what you usually do each day						

	agency	heel	pavement	penalty	principal	youth
back of your foot						
person in charge of a school						
punishment						

	element	jail	joint	objective	portrait	variety
goal						
picture						
place where criminals are kept						

	defeat	infant	nuclear	outrage	prospect	rival
loss						
person you oppose						
small child						

	coincide	derive	devote	permit	publish	regret
feel bad about doing something						
give all your time and attention						
happen at the same time						

	civilize	discharge	graduate	imply	merge	perceive
join						
release						
suggest						

	assault	bargain	compete	dedicate	nominate	restrain
attack						
hold back						
try to win						

	fundamental	humorous	interior	numerous	prompt	religious
basic						
many						
on time						

	legislative	mechanic	mortal	random	rear	reluctant
back of something						
can die						
without order						

4,000 Word Level

	auction	bullet	fever	flock	outlet	skull
group of birds						
high body temperature						
sale where people place bids						

	archive	ash	mat	moisture	physics	tile
place where old books are kept						
powder left after something burns						
science subject						

	pioneer	dictionary	immigration	petition	romance	thigh
book with information given for each word						
first person to do something						
paper that people sign						

	acid	cafe	deadline	deficiency	texture	thesis
lack						
place for buying and drinking coffee						
time limit						

	avenue	brass	departure	hood	hut	premier
cover for your head						
small house						
type of metal						

	appall	invade	mutter	refine	roast	unveil
cook over fire						
enter by force						
make pure						

	aspire	exert	gossip	minimize	poke	postpone
make smaller						
push with your finger						
try to reach a goal						

	adhere	fracture	originate	peel	sparkle	terminate
do what is expected						
end						
give off small flashes of light						

	amateur	arrogant	cognitive	infinite	judicial	monetary
having no limits						
not professional						
overly proud						

	delicate	dull	miserable	noble	peculiar	refreshing
breaks easily						
unselfish and morally good						
very unhappy						

5,000 Word Level

	calf	epidemic	foam	landmark	token	trumpet
illness spread quickly that affects many people						
many bubbles						
young cow						

	comb	**ivory**	**pants**	**rainbow**	**vegetarian**	**zip**
containing no meat						
hard white substance						
tool for styling hair						

	analogy	**captive**	**remainder**	**renovation**	**ribbon**	**vest**
comparison between two things						
person kept somewhere unwillingly						
what is left						

	butcher	**chalk**	**grape**	**ornament**	**pier**	**wallet**
container for money						
person who cuts and sells meat						
place for boats to dock						

	ammunition	**crab**	**dusk**	**nucleus**	**revenge**	**spectator**
beginning of night						
center						
person who watches						

	abolish	**apprehend**	**chuckle**	**erode**	**replicate**	**segregate**
end						
keep apart						
slowly make smaller						

	duplicate	**emigrate**	**hurl**	**perch**	**revolt**	**swirl**
copy						
fight violently against						
sit in a high place						

	amplify	**evaporate**	**grunt**	**mitigate**	**recollect**	**tow**
disappear						
make larger						
remember						

	blunt	fabulous	horrified	numb	singular	volatile
not sharp						
without feeling						
wonderful						

	brisk	extinct	fragrant	splendid	tolerant	trivial
fast						
having no living members						
of little importance						

Authors' addresses

Stuart Webb
Faculty of Education
University of Western Ontario
1137 Western Road, London
Ontario, N6G 1G7
Canada

swebb27@uwo.ca

Yosuke Sasao
Institute for Liberal Arts and Sciences
Kyoto University
Yoshida Nihonmatsu-cho, Sakyo-ku
Kyoto 606–8501
Japan

sasao.yosuke.8n@kyoto-u.ac.jp

Oliver Ballance
Linguistics and Applied Language Studies
Victoria University of Wellington
Office 201, 22 Kelburn Parade
Wellington, 6012
New Zealand

Oliver.Ballance@vuw.ac.nz