CREATING AND TRIALLING SIX VERSIONS OF THE VOCABULARY SIZE TEST

Averil Coxhead, Paul Nation and Dalice Sim
Victoria University of Wellington

Abstract

This paper reports on research in progress on the creation and trialling of six versions of a 20,000 version of the Vocabulary Size Test (VST) (Nation & Beglar, 2007). The VST is beginning to be used by language teachers in various contexts. Six 100 item versions of the VST were developed and trialled with 46 test-takers who sat all six versions of the test. The results indicated that there were two sets of three parallel versions with generally no statistical differences between each set. The equivalence of the tests was checked across a range of variables including first language, gender, status as a university student, age and level of education. The paper suggests limitations of the VST, some cautious implications, and further research.

Introduction

Knowing the vocabulary size of language learners is important for setting goals in a classroom programme (Nation & Webb, 2011) and possibly providing a diagnosis for learners who have problems with reading and writing (Nguyen & Nation, 2011, p. 87). The Vocabulary Size Test (VST) (Nation & Beglar, 2007) was designed to measure both first and second language learners' written receptive vocabulary size in English. The test measures knowledge of written word form, the form-meaning connection, and to a smaller degree concept knowledge at the item level. Analysed using Read and Chapelle's (2001) framework, the VST is a discrete, selective, relatively context-independent vocabulary test. At the test level, it provides a rough estimate of total vocabulary size where vocabulary knowledge is considered as including only single words (not multiword units) and vocabulary size does not include proper nouns, transparent compounds, marginal words like *um, er, gee gosh,* and abbreviations. The VST does not measure the ability to distinguish homonyms and homographs.

The original version of the VST tested up to the 14,000 level and was developed by Paul Nation. This means the test starts by testing words at the 1,000 frequency level, then the 2,000, then the 3,000 and so on up to 14,000. This version has 140 multiple-choice items, with 10 items from each 1000 word family level from the most frequent 14,000 word families of English. A learner's total score needs to be multiplied by 100 to get their total receptive vocabulary size (Nation & Beglar, 2007). Test-takers select the best definition of each word from four choices. Here is an example item:

16. strap: He broke the <strap>.
    a. promise
    b. top cover
    c. shallow dish for food
    d. strip of strong material

Beglar's (2010) examination of the 140 item VST showed it can be used with learners with a very wide range of proficiency levels and it clearly measures a single factor (written receptive vocabulary knowledge). He also found the test has a range of item difficulties related to the frequency level of the tested words. Beglar compared the performance of male participants with female participants, versions of the test with different numbers of items, and learners of various proficiency levels. Rasch reliability measures were around .96.

Nguyen and Nation (2011) showed that it is important to sit all levels of the test because some words at the lower frequency levels will be known. This may be because they are loan words or cognates, they relate to learners' hobbies and interests, they are technical words in fields the learners are familiar with, or the learners just happened to meet and learn them.

The issue of cognates is important in the VST. Removing the loanwords or cognates in the learner's first language from the test would distort the measurement of vocabulary size because they are a legitimate part of a learner's second language vocabulary size. Because cognates are so influential, it may be necessary to ensure that when the test is used with learners with the same first language, the proportion of cognates in the test reflects the proportion of cognates in the language (Elgort, 2013; see also Elgort & Coxhead, in press).

The 20,000 word family versions of the VST were developed to reduce the ceiling effects of the 14,000 level test and because the test should measure frequency levels beyond the test-takers' likely vocabulary size. The larger version means that the test could be used with adult native-speakers as well as high proficiency non-native speakers. For more on the vocabulary size testing research from Victoria University of Wellington, see Nation and Coxhead (2014).

The advantages of having parallel versions of a test

One practical advantage of parallel versions is test security. It lessens the chance that a learner who has just sat the test can inform others who are yet to sit the test. It also reduces the effect of an earlier test on a later test. Having parallel versions also makes longitudinal research on vocabulary size and growth much more manageable because the same version of a test does not have to be used over and over again. Finally, comparing the same learners' results on parallel versions of a test is also a way of assessing the reliability of a test. Carmines and Zeller (1970, p. 40) note that,

> The alternative-form method for assessing reliability is obviously superior to the simple retest method, primarily because it reduces the extent to which individuals' memory can inflate the reliability estimate.

The basic limitation of the alternative-form/parallel-form method is the practical difficulty of constructing alternative versions that are parallel. The Vocabulary Levels

Test (Nation, 1993; Schmitt, Schmitt & Clapham, 2001) is an example of a test which has parallel versions.

Criticisms of the VST

The multiple-choice format provides opportunities for guessing, which might be done by elimination of choices. Because each item represents 200 word families, random guessing can inflate scores. The amount of random guessing will depend on the way the test is administered (one-on-one versus group administration), learners' attitudes to the test, and learners' vocabulary size (learners with larger vocabulary sizes have fewer items that they truly don't know). When interpreting the results of the test, it needs to be remembered that multiple-choice tests (recognition tests) give higher scores than translation or interview tests (recall tests) (Laufer & Goldstein, 2004). Recall tests tend to underestimate vocabulary knowledge, while recognition tests overestimate vocabulary knowledge.

The VST is designed to give credit for partial knowledge because the distractors are not closely related in meaning to the correct choice (Nagy, Herman, & Anderson, 1985). Partial knowledge may be sufficient to cope with a word and learn more about it, when it is met in context while reading. Another criticism of the test might be that it is based on word families and there is no guarantee that knowing a headword of a word family implies knowledge of the other words in a family. Keeping these criticisms in mind, this paper reports on an analysis of the six versions of the VST, through an analysis of the results from 46 participants.

Research questions

1. To what extent can the six versions of the test be considered parallel or equivalent across the 46 individual test takers?
2. What effect might the test order have on the test results?
3. What effect might first language, gender, current university study, age, and level of education have on the results of the tests?

Methodology

Developing six versions of the VST

The words in the Vocabulary Size Test were sampled from word family lists originally created from data from the British National Corpus (Nation, 2006). Sampling from frequency-based word lists avoids the severe sampling biases that occur when sampling from dictionaries (Nation, 1993). Distractors were definitions of words chosen from the same 1000 word level as the tested word. Care was taken with the length of the options. Finally each test was run through the Range program (Heatley, Nation & Coxhead, 2002) to double check that distractors were the same level as their test items. This procedure was used for five versions (B-F). Version A consists of the original 14,000 version with six new levels added (15,000-20,000), using the procedure as for the other versions.

Item sampling and test format

Each test contains a total of 100 items, five from each of twenty 1,000 word level bands. The margin of error of a sample is primarily determined by the size of the sample not the sampling rate. For the VST, 100 items reliably represent the combined 20 bands because 100 items are a large enough sample to represent 20,000 words. However, the test cannot be used to see what proportion of words is known at each 1000 word family band because five items are not enough to represent a band. During initial trialling, some of these tests were combined into 200 item versions, but Beglar's (2010) findings and our own piloting showed that a 100 item test could be sat in a reasonably short period of time, and that the scores were consistent with 100 and 200 item tests (Coxhead, unreported data). Because the sampling rate from the BNC lists is one in 200, scores on the 100 item versions of the VST need to be multiplied by 200 to estimate total vocabulary size.

Participants

Almost all the 46 participants were university students who ranged in age from 16 years old to two people over 60, with most in their twenties. 28 were native speakers of English and the rest were high proficiency non-native speakers. There were 34 females and 12 males. The participants are a convenience sample. It was difficult to find test-takers who were willing and could spare the time (around three hours) to sit six tests.

Administration of the tests

The tests were administered on computer with a researcher present. Computer scoring ensured reliability of scoring. About one third of the test-takers sat all six tests at once with rest breaks, and the remainder sat three tests at a time in two sessions. Each test took between 20 to 40 minutes. The participants sat the tests individually, not in groups.

Results and discussion

Research question one: To what extent can the six versions of the test be considered parallel or equivalent across the 46 individual test takers?

To decide if the six tests were equivalent (parallel versions), we needed to see if the mean scores and variances on the different versions were significantly different from each other, and how much an individual's score would differ when sitting two different versions. We compared each test to all other tests using the methods of Bland and Altman (1986). Here equivalence of two tests is assessed using the mean difference ($\bar{d}$), and its standard deviation, s. If differences within $\bar{d} \pm 2s$ would not lead to differences in interpretation of the result, then the two tests could be used interchangeably. Table 1 provides the descriptive statistics for the six tests, ordered by mean score.

Table 1:
Descriptive statistics for the six versions of the Vocabulary Size Test

Test version	**Mean**	**Std. Deviation**	**N**
B	83.20	13.982	46
A	81.37	16.662	46
D	81.33	14.592	46
C	78.74	15.221	46
F	78.65	14.439	46
E	78.20	13.341	46

The first question was whether the test results (total number correct out of 100) varied significantly by test (A – F). All the participants took all 6 tests, so repeated measures analysis of variance was used to compare the mean results, while taking into account the correlated responses (that is, the same person's results on two different tests would be expected to be correlated). To compare different subgroups of participants (e.g., native and non-native speakers, see research question 3 below), the subgroup was used as a between group factor and test as a 'within' factor in the repeated measures design.

Since, by Mauchly's test, we cannot assume sphericity with these data, the results of the Greenhouse-Geisser statistic was used because it adjusts for lack of sphericity in the data to test the overall hypothesis that the results vary by test. We concluded ($F(4.061, 182.759) = 14.573$, $p < 0.0005$) that the mean test results differed significantly by test.

We then used the Bonferroni correction for the pairwise comparisons to see which tests were different from which other tests. From these results, we can conclude that tests A, B and D have significantly ($p < 0.05$) higher mean results than tests C, E and F. We cannot statistically differentiate between A, B and D, nor between C, E and F. Figure 1 is the plot of the predicted mean results. Note that the scale on the vertical axis covers a small range of scores.

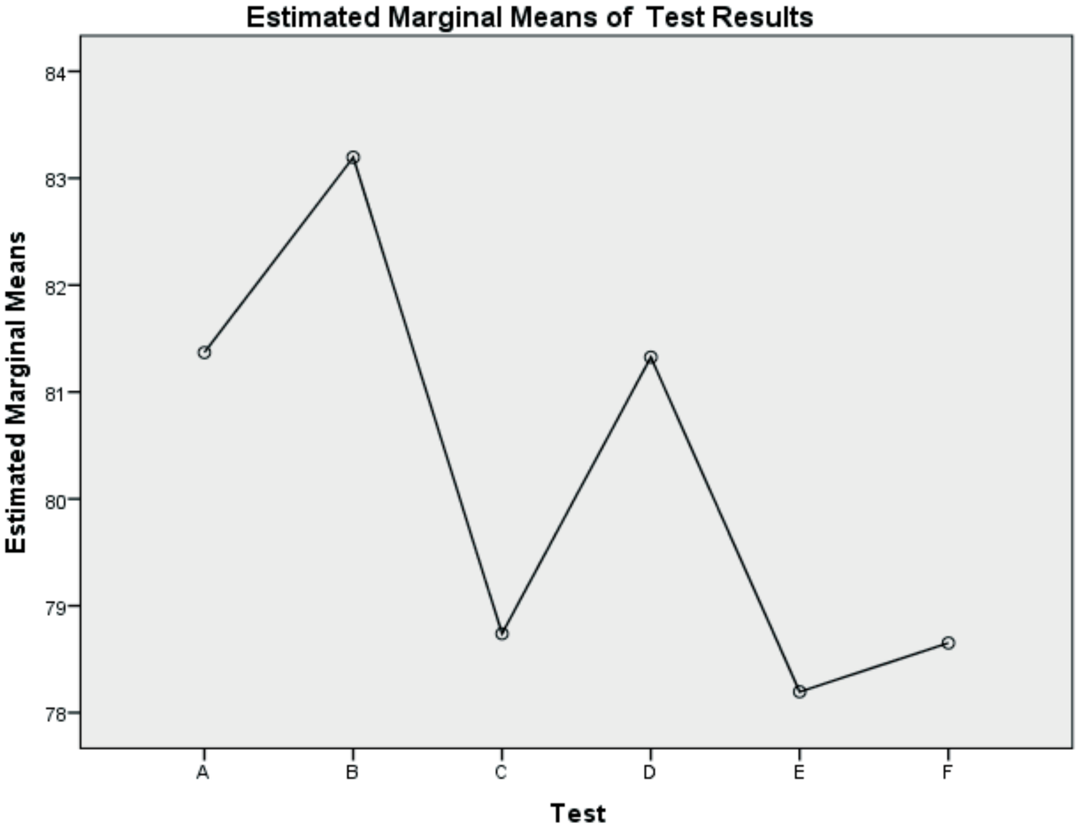

Figure 1:
Means of six versions of the Vocabulary Size Test

In measures of total vocabulary size, we would be satisfied if two or more tests placed the same person roughly within the same 1000 level band. If scores for the same person differed by less than 5 out of 100, that would be satisfactory equivalence. The highest mean score (Test B: 83.20) and the lowest mean score (Test E: 78.20) differ by five points, which is on the margins of being too far apart. Within each grouping of three tests, (A, B, D) (C, E, F) the mean differences between the highest and lowest are much smaller (1.87, and 0.54).

Table 2:
Mean differences and 95% confidence intervals for the six test versions

Comparison	**Mean Diff**	**SD Diff**	**Lo Limit 95%**	**Hi Limit 95%**
A vs B	-1.8261	4.29110	-3.07	-.59
A vs C	2.6304	5.11798	1.15	4.11
A vs D	.0435	4.29965	-1.20	1.29
A vs E	3.1739	6.05400	1.42	4.92
A vs F	2.7174	5.98392	.99	4.45
B vs C	4.4565	4.88402	3.05	5.87
B vs D	1.8696	4.42020	.59	3.15

B vs E	5.0000	4.47710	3.71	6.29
B vs F	4.5435	5.82602	2.86	6.23
C vs D	-2.5870	4.84688	-3.99	-1.19
C vs E	.5435	4.34975	-.71	1.80
C vs F	.0870	4.95682	-1.35	1.52
D vs E	3.1304	5.22647	1.62	4.64
D vs F	2.6739	4.97127	1.24	4.11
E vs F	-.4565	5.44755	-2.03	1.12

From Table 2, we see that the 95% confidence intervals for the mean differences are no more than 4 points (in either direction) for comparisons between A, B and D. The same is true for comparisons between C, E and F. Comparing between these two groups of three test versions, the 95% confidence limits are almost all 4 (A vs C, A vs E, A vs F, D vs E, D vs F) or more (B vs C, B vs E, B vs F). For assessing vocabulary size, these data support the conclusion from the repeated measures analysis of variance that the six test versions fall into two groups of three equivalent versions: (A, B, D) and (C, E, F).

If two tests are equivalent, they give the same information about the test participants, so we would expect the tests to have the same or similar variances in the same sample of participants. This would indicate that both tests find the same spread of responses within the same sample. We did Fisher's test to compare variances between each pair of tests (Snedecor & Cochran, 1980, pp. 98-99). None of the variances was statistically different from any other, the smallest p-value being 0.0699 for the comparison of A with E. Therefore, comparing the variances of the six tests provides no evidence to support or refute the two sets of three parallel tests.

It is important to know if different tests give the same score or close to the same score for each of the individuals for each pair of comparisons (15 comparisons, A vs B, A vs C, A vs D, A vs E, A vs F, B vs C, and so on) between the six versions of the tests. Looking at this involved a total of 46 x 15= 690 comparisons. The aim was to see how many of the 690 comparisons were identical scores, or differed by 1 point out of 100, 2 points, 3 points and so on. The comparisons were done for each of the two groups of three tests, and for all six tests. If most of the comparisons were identical scores or within five or less points of each other, this would give us greater confidence in using the tests as parallel versions for looking at individuals.

Table 3:
Percentage of participants with pairwise differences within 3 or 5 points for the parallel versions

Comparison	**Percent 0**	**Percent < 3**	**Percent < 5**
A vs B	15.2	60.9	73.9
A vs D	10.9	67.4	80.4
B vs D	6.5	63.0	87.0

C vs E	2.2	56.5	73.9
C vs F	8.7	41.3	69.6
E vs F	6.5	52.2	71.7
Average	8.3	56.9	76.1

The percentages in each row in Table 3 are cumulative. Table 3 shows that 15.2% of the 46 test-takers got exactly the same score on Tests A and B, 60.9% got identical scores or scores differing by 3 or less on Tests A and B, and 73.9% of the test takers got scores differing by 5 or less. On average, just over 76% of the test-takers got scores within 5 points of each other on the parallel versions. However, approximately one quarter of the test-takers had scores differing by six points or more.

Table 3 also shows that tests A, B and D (the first three comparisons) are very consistent, with a high percentage of participants (73.9% - 87.0%) scoring within 5 points on these three tests. Tests C, E and F are less consistent (the second three comparisons) with 69.6% - 73.9% of participants scoring within 5 points on these three tests. Not all of these differences will be the fault of the tests themselves, because the differences can also come from the learners and the care with which they sat the tests. Test-retest data on exactly the same versions is needed to act as a comparison control for these variables.

Let's now look at pairs of tests between the two groups of ABD and CEF. So, A is compared with C for each person, A with E and so on (nine comparisons). See Table 4 below.

Table 4:
Pairwise comparisons of differences in test scores for each test-taker expressed in percentages of test-takers for the non-parallel versions

Comparison	**Percent 0**	**Percent < 3**	**Percent < 5**
A vs C	4.3	45.7	60.9
A vs E	4.3	34.8	54.3
A vs F	4.3	32.6	63.0
B vs C	4.3	45.7	58.7
B vs E	2.2	23.1	56.5
B vs F	2.2	39.1	58.7
D vs C	15.2	43.5	69.6
D vs E	8.7	41.3	54.3
D vs F	6.5	45.7	69.6
Average	5.8	39.1	60.6

If we compare the averages in the bottom row of Table 3 (the two sets of three parallel versions comparisons) with those in Table 4 (the nine non-parallel versions comparisons), we see a greater likelihood of closer scores when sitting two parallel versions. 60.6% of the test-takers had scores within five points or less of each other on the non-parallel versions compared with 76.1% of the test-takers on the parallel versions.

This data supports the two groupings of parallel versions but shows very clearly that even parallel versions are unlikely to provide identical scores, and for a significant group of test-takers (around 20-30%) the scores on two tests taken by the same person are likely to be several points apart.

Research question two: What effect might the test order have on the test results?

The order of the tests was varied for different learners so there was no one set order for taking the six tests. However, scores on tests sat later could have benefited from the test-takers' experience of sitting the previous tests. Alternatively, tests sat later could have been affected by test-taking fatigue. The overall means and standard deviations of test results by order are shown in Table 5. In column 1 of Table 5, the number 1 refers to tests sat first, 2 to tests sat 2nd and so on. If order of sitting has an effect on the results, we would expect to see either a rise in mean scores as we move down Table 5 as a result of improvements in test-taking skill through practice, or a drop as a result of fatigue or declining commitment. There is no evidence of such changes.

Table 5:
Order, means and standard deviations

Order	**Mean (sd)**	**n**
1	80.54 (15.49)	48
2	83.26 (13.66)	47
3	79.55 (15.20)	49
4	78.55 (14.21)	42
5	77.31 (15.89)	55
6	83.43 (12.55)	35
Total	80.25 (14.72)	276

The number of people in column 3 is sometimes higher than 46 because some tests were sat in 200 item versions, and they were counted as being sat simultaneously. So, two people sat 200 item versions as their first test and the two 100 tests in this 200 version were both counted as being sat first, raising the n from 46 to 48.

Research question three: What effect might first language, gender, current university study, age, and level of education have on the results of the tests?

We next wanted to see whether or not this difference between the tests was maintained in different subgroups of the participants. The subgroups were native speakers (n=28) and non-native speakers (n=18), gender (male n=12; female n=34), status as a university student (n=31) versus not studying (n=15), different age levels, highest level of education, and first language group (English - n=28; European - n=7; Other - n=11). Table 6 below contains the results divided into these subgroups. The order of the tests in column 1 of Table 6 is based on the ranked mean scores from Table 1, so B was the test with the highest mean and E was the lowest.

Table 6:
Means and standard deviations for the six tests comparing native speakers and non-native speakers, females and males, and status as a university student

Test	**Native**	**Non-native**	**Female**	**Male**	**Studying**	**Not studying**
B	90.79	71.39	85.67	76.92	80.00	89.80
(sd)	(5.61)	(15.00)	(13.05)	(14.83)	(14.55)	(10.28)
A	89.93	68.06	84.79	72.69	77.35	89.67
(sd)	(6.90)	(18.73)	(15.18)	(17.68)	(17.30)	(11.90)
D	88.96	69.44	84.33	73.69	77.71	88.80
(sd)	(6.60)	(15.78)	(12.91)	(16.31)	(15.46)	(9.12)
C	86.64	66.44	81.27	72.31	74.68	87.13
(sd)	(7.29)	(16.31)	(14.87)	(14.72)	(15.88)	(9.61)
F	86.18	66.94	81.64	71.08	74.68	86.87
(sd)	(6.01)	(16.00)	(13.54)	(14.36)	(14.81)	(9.65)
E	84.96	67.67	80.55	72.23	74.48	85.87
(sd)	(6.39)	(14.61)	(13.19)	(12.42)	(13.84)	(8.26)

The table shows that scores within each group of three tests are very close to each other. Repeated measures ANOVAs showed no significant difference within each of the two groups of three. Mauchly's test of sphericity was significant for all comparisons, and so the Greenhouse-Geisser statistic was used, and showed that both groups in each comparison followed the same pattern in their test scores. For all the people tested, native speakers predictably scored higher than non-native speakers. Table 6 shows large standard deviations for the non-native speakers indicating a wide range of English proficiency levels. Females had larger vocabulary sizes than males and those not currently studying at university had higher scores than those who were studying at university. Because the people tested were not randomly chosen and are unlikely to be representative of the general population, not too much can be generalized from these scores.

Comparisons of the means on the six tests across age groups, level of education and first language generally supported the groupings of the two sets of three tests. Table 7 shows the effect of first language on performance on the test. The speakers of European languages which have cognate relations with English were in one group (n=7). This included speakers of German, Italian, Dutch and Afrikaans. Speakers of other languages without cognate relations to English (n=11) were in another group and included speakers of Indonesian, Malay, Vietnamese and Chinese.

Table 7:
Effect of first language on mean scores and standard error of the six versions

Test	**English (n=28)**	**European (n=7)**	**Other (n=11)**
	Mean (sd)	**Mean (sd)**	**Mean (sd)**
B	90.79 (1.35)	86.14 (2.69)	62.00 (2.15)
A	89.93 (1.55)	87.57 (3.10)	**55.64** (2.47)
D	88.96 (1.42)	85.57 (2.84)	59.18 (2.27)
C	86.64 (1.57)	82.57 (3.14)	56.18 (2.51)
F	86.18 (1.44)	82.71 (2.87)	56.91 (2.29)
E	84.96 (1.48)	81.14 (2.97)	**59.09** (2.37)

A ceiling effect is operating for both native-speakers (note the low standard errors) and speakers of European languages. Table 7 illustrates how close the first language speakers of European languages are to the scores of the native English speakers. The two groupings of three tests are once again maintained in the European language speakers' scores. Only the score for European A is out of order within that group. The group of speakers of other languages do not follow the previous patterns so well (see the bolded results in column 4 of Table 7). The score for A is lower than it should be, and the score for E is higher than it should be, perhaps a result of cognates or loan words.

The statistical tests from the repeated measures ANOVA verify these results. The correlations between the test results do not meet the criterion for sphericity, so the Greenhouse-Geisser adjustment was used. Overall, there was a significant difference between tests, $F(3.952, 169.932) = 10.465$, $p < 0.0005$. The pattern of results across tests was different for the different language groups, $F(7.904, 169.932) = 3.034$, $p = 0.003$. That is, the plots of results by test are not parallel for the three language groups. Averaged over all six tests, there was a difference in score by language group, $F(2,43) = 70.143$, $p < 0.0005$. Multiple comparisons indicate that group 3 (Other) is significantly different from groups 1 and 2, which are statistically similar. Learners of English as a foreign language who are native speakers of a European language typically achieve much higher proficiency than native speakers of other languages.

The plots are:

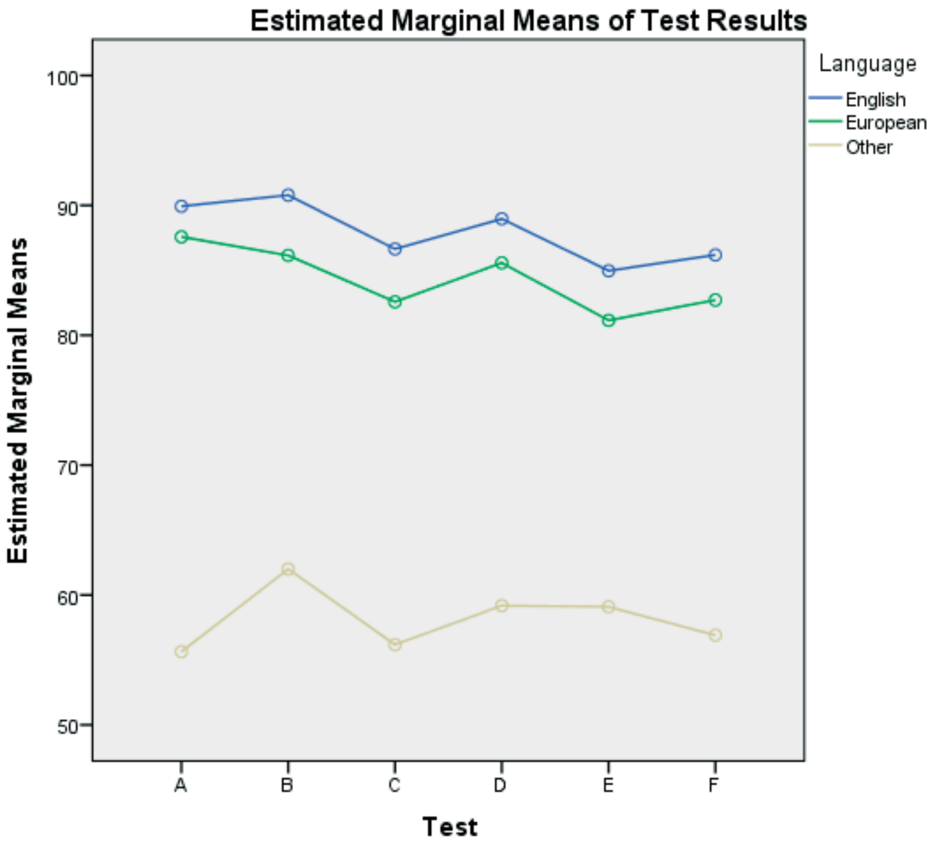

Figure 2:
Means on the six versions of the Vocabulary Size Test for native speakers of English, speakers of European languages, and speakers of other languages

Those with higher education had higher scores (PhD 1, High school 1, degree 44) and, generally, older test-takers had higher scores than younger test-takers. However, some of the groupings had very small numbers, for example, only 2 people in the 61+ age group and 5 in the 16-20 age group, so we cannot depend on these findings.

Based on our data, we can conclude that tests A, B and D score consistently higher than tests C, E and F although this is on average with a difference of a few points out of 100. This pattern was repeated in all subgroups, except for the first language subgroups, suggesting that the differences seen between tests will be consistent across most subgroups. The choice of test to be used may, however, potentially advantage or disadvantage different groups of participants according to their first language, A and E being out of order (Table 4, column 4). Although the difference between the two sets of three tests is rather small, researchers using two or three versions should use the tests within one group of three.

To work out what a score on the tests means in terms of language use, we need to look at the vocabulary size needed to gain a text coverage of 98% in various kinds of texts. Table 8 provides such data.

Table 8:
Vocabulary sizes needed to get 98% coverage (including proper nouns) of various kinds of texts (Nation, 2006)

Texts	**98% coverage**	**Proper nouns**
Novels (Nation, 2006)	9,000 word families	1-2%
Newspapers (Nation, 2006)	8,000 word families	5-6%
Children's movies (Nation, 2006)	6,000 word families	1.5%
General Spoken English (Nation, 2006)	7,000 word families	1.3%
Spoken Academic English (Dang & Webb, 2014)	4,000 word families plus proper nouns and marginal words	0.37-1.69%
TED Talks (Coxhead & Walls, 2012)	9,000 word families plus proper nouns (1.44%)	1.44%)

Note that Dang & Webb (2014) report 96.05% coverage over the British Academic Spoken English corpus at 4,000 word families plus proper nouns and marginal words, and 98.00% coverage at 8,000 word families plus proper nouns. Note that the range of proper nouns in this study differs across academic disciplines. Coxhead & Walls (2012) found the vocabulary load of 9,000 plus proper nouns over a corpus of TED Talks.

The goal of around 8,000-9,000 word families is an important one for learners who wish to deal with a range of unsimplified spoken and written texts. It is helpful to know how close learners are to this critical goal. Initial studies using the test indicate that undergraduate non-native speakers of non-European backgrounds successfully coping with study at an English speaking university have a vocabulary size around 5,000-6,000 word families. Non-native speaking PhD students have around a 9,000 word vocabulary.

Limitations

The small number of participants is an important limitation. It is a matter of debate whether a cut-down version of the test, for example a 50 item test going up to the 10th 1000 words, is better for intermediate learners of English as a foreign language. Limiting the size of the test like this will have the negative effect of not allowing the learners to show knowledge of the low frequency words that they happen to know. The positive effects will be to reduce the time to sit the test, the elimination of a large number of items that learners do not know, and the subsequent reduction of the effect of random guessing.

Note that an item analysis for the new versions is needed to establish whether all of the items in the new versions are working correctly. These results will be reported in another paper.

Implications, future research, and conclusion

This research suggests there are two sets within the six versions of the VST, and that we need to be cautious with releasing these versions until further validation work has been carried out. It is clear from the data that tests (B, A, D) have equivalent means and variances, and are more likely to provide roughly similar scores. C, F, and E can also have equivalent means, but they are not as "consistent" as the ABD grouping, in that a lower percent of participants had scores fewer than 5 points apart. All six tests give average scores within 5 points of each other which would roughly place test-takers within the same 1000 level band. In order to achieve greater reliability, it may be wise to follow Diack's (1975) guidelines. That is, to get test-takers to sit more than one test and calculate the average.

Future work on the VST will include more validation research, investigating the VST alongside other measures of vocabulary development (Elgort & Coxhead, in press for more), and a VST for listening is under development by researchers in Japan. More work also needs to be done on the effect of test-taking strategies and the VST, to probe the problem of guessing with the multiple choice format (Elgort & Coxhead, in press).

Teachers need to be cautious with administering the test to large groups because of the need for engagement with the test. In this study, participants sat the tests in one-on-one conditions with a researcher, which ensured that the test-takers remained focused on the task. We also caution against using the larger versions of the test with beginner or even intermediate level learners of English, considering the possible motivational impact of encountering words in a test which even adult native speakers of the language might not know. Students (native and non-native speakers) might want to test their vocabulary independently or teachers might want to administer the test to classes at the beginning of a course to help set or modify learning goals. As more bilingual versions of the test become available, the VST will be able to be used to find out more about the vocabulary size of lower proficiency learners of different first languages. Finally, it is important to resist the urge to consider this test in any way a levels test for the BNC 1000 lists. Versions A and B are on Paul Nation's website (http://www.victoria.ac.nz/lals/about/staff/paul-nation). Two versions will be kept in-house and the final two versions can be obtained for research purposes by contacting the authors.

Acknowledgements

This research was supported by a grant from the Faculty of Humanities and Social Sciences, Victoria University of Wellington. Many thanks to Alex Heatley for his IT support, Brian King for test development, and our highly skilled research assistants. Finally, thank you to all the participants.

References

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

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical assessment. *Lancet, i*, 307-310.

Carmines, E. G., & Zeller, R. A. (1970). Reliability and validity assessment. *Sage University Paper, 17*.

Coxhead, A. & Walls, R. (2012). TED Talks, vocabulary, and listening for EAP. *TESOLANZ Journal*, *20*: 55-67.

Dang, Y. & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes,* 33: 66–76.

Diack, H. (1975). *Test your own wordpower*. St. Albans: Paladin.

Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning, 61*(2), 367-413.

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

Elgort, I., & Coxhead, A. (in press). An introduction to the Vocabulary Size Test: Description, application and evaluation. In J. Fox & V. Aryadoust (Eds.) *Current trends in language testing in the Pacific Rim and the Middle East: Policies, analysis, and diagnosis*. Cambridge: Cambridge Scholars.

Heatley, A., Nation, P., & Coxhead, A. (2001). The RANGE Programme. Available at http://www.victoria.ac.nz/lals/about/staff/paul-nation.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning, 54*(3), 399-436.

Nagy, W. E., Herman, P., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly, 20*(2), 233-253.

Nation, I. S. P. (1993). Using dictionaries to estimate vocabulary size: Essential, but rarely followed, procedures. *Language Testing, 10*(1), 27-40.

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

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

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

Nation, P. and Coxhead, A. (2014). Vocabulary size research at Victoria University of Wellington, New Zealand. *Language Teaching*, 47(3): 398 – 403.

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.

Read, J., & Chapelle, C. (2001). A framework for second language vocabulary assessment. *Language Testing, 18*(1), 3-32.

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.

Snedecor, G. W., & Cochran, W. G. (1980). *Statistical Methods* (7th ed.). Ames, Iowa: Iowa State University Press.