An Examination of the Behaviour of Four Vocabulary Tests

Norbert Schmitt, University of Nottingham

There has been a continuing interest in measuring subjects' total vocabulary from at least the early part of this century (Cuff, 1930) right up to the present time (Meara, 1992). Current belief in the value of vocabulary measurement is witnessed by the fact that many commercial second language proficiency tests, such as the TOEFL test (Educational Testing Service, 1987), include a vocabulary component. Also, many, if not most, second language classroom teachers write and give vocabulary tests of one kind or another to their students (see Schmitt, 1994 for a framework for writing vocabulary tests). Yet, in spite of the widespread use of vocabulary tests, the area of vocabulary testing has not attracted a proportionate amount of research attention. The fact that there is no generally accepted vocabulary test which can be used as a standard illustrates this deficit in research, as well as hindering future research in the field.

This paper takes advantage of the data generated from a larger study to look at four different vocabulary tests. The analysis will examine the tests from three perspectives: first the tests will be correlated against each other and the TOEFL test, next they will be correlated to two independent measures, and finally the scores of individual subjects will be examined to see if this illuminates the behavior of the various tests.

A Brief Description of the Vocabulary Tests

The Vocabulary Levels Test **[Levels Test]** (Nation, 1983)

The Levels Test attempts to measure a subject's vocabulary at a number of stages in the word frequency range, namely the 2,000, 3,000, 5,000, and 10,000 word levels, as well as including the 800 words in the University Word List (Nation, 1990, p. 235-239). The test uses a matching format.

Written Checklist Test **[Checklist Test]**

Checklist tests simply ask subjects to tick words which they believe they know. These tests also contain a number of nonwords, which if ticked, reduce the total score according to a predetermined formula. See Meara (1992) and Meara and Buxton (1987) for a more detailed description.

The Eurocentres' 10K Vocabulary Size Test **[EVST]**

(Meara and Jones, 1990)

The EVST is a computerized version of a checklist test. Target words appear on the screen and the subjects must decide whether they know them and push the Y (Yes) or N (No) keys.

The Test of English as a Foreign Language **[TOEFL]**

(Educational Testing Service, 1987)

This widely-used second language proficiency test consists of three sections: language structure and writing, listening, and vocabulary and reading. Although the last section is not wholly a vocabulary test, it will also be included in the analysis as **TOEFL V/R**.

The 8-month longitudinal study of Japanese students from which the data has been extracted (Schmitt, n.d.) primarily examined how native-like the subjects were in their knowledge of word associations and derivational suffixes. Thus, the resulting two measures give an indication of *how well* the subjects knew each of the target words. If it can be accepted that learners who have larger vocabularies also tend to have more 'in-depth' knowledge of the individual words, then these two measures of component word knowledge can provide a benchmark from which to judge the above vocabulary size tests.

Comparison of the Vocabulary Tests with Each Other and the TOEFL Test

The aforementioned lack of a standard vocabulary test makes it impossible to judge vocabulary tests against any accepted benchmark. It thus makes it necessary to use less direct methods of evaluation. The first used in this study involves running correlations among the four vocabulary tests and the TOEFL test to discover to what extent they are measuring the same thing. The results appear in Table 1 for two tests, one (T1) at the beginning of the school year and one (T2) at the end.

Table 1: Correlations of Vocabulary and TOEFL Tests :

Test 1 (T1)[1] N=26	Checklist	Levels	TOEFL	TOEFL Voc/Reading
EVST	.433	.487	.581	.611
Checklist	---	.316	.399	.524
Levels		---	.686	.580
Test 2 (T2)[2] N=21				
EVST	.654	.766	.449	NS
Checklist	---	.273	NS	NS
Levels		---	.625	.590

[1] Beginning of school year

[2] End of school year

All correlations r<.05

The first notable point is that the written checklist test seems generally to be the 'outsider'. It had the weakest correlations with the other tests and either correlated only weakly or not at all with the TOEFL proficiency test. This can be largely explained by the brevity of the test. It was written with only 50 items in order to determine whether a short easy-to-give-and-correct written checklist test would yield similar results to longer, more involved tests. The answer appears to be negative. This is not a fault of the checklist format, however, as it seems that if a series of three or four checklist tests are given (totaling perhaps 150 words), the combined result gives a reasonable, reliable estimate of vocabulary size (Meara, personal communication).

In contrast, the Levels Test and the EVST seemed to be measuring more of the same underlying vocabulary knowledge. On the T2, they correlated at .76. Unfortunately, this is not nearly as high as one would expect, especially since they both attempt to measure precisely the same thing. In fact, if that correlation is squared to produce a figure for covariance, the result is only about .58. It is a little disheartening that two of the most advanced vocabulary tests we have at the moment failed to correlate more closely.

The vocabulary test results also correlated significantly with the TOEFL test results. In this case, the Levels Test had the strongest correlations (.62 and .68). If we believe that the tests are giving reasonably accurate estimates of vocabulary size and general language proficiency respectively, then these correlations give yet more evidence of the important relationship between the two.

The fact that the EVST and Levels Test often correlated more strongly with the overall TOEFL scores than with the TOEFL V/R scores came as a surprise. Three possible explanations suggest themselves. First, the reading component might have changed the scores quite radically from what they would have been if only the vocabulary test were included. Second, since vocabulary is so critical to general language proficiency, vocabulary test results might correlate more strongly with general language proficiency measures than they do with a combination of vocabulary and reading measures. Third, the vocabulary component of the TOEFL itself could be suspect, not giving very good estimates of vocabulary size. Unfortunately, there is not enough information to make a principled choice from among these explanations.

Comparison of the Vocabulary Tests with Measures of Association and Suffix Knowledge

The next stage of the analysis involved correlating the test scores against the measures of word association and verbal suffix knowledge. The results are illustrated in Table 2.

Table 2: **Correlations of Vocabulary Tests and Suffix and Association Measures T1 N=26; T2 N=21**

	Productive				Receptive			
	Suffix		Association		Suffix		Association	
	T1	T2	T1	T2	T1	T2	T1	T2
EVST	NS	NS	NS	.52	NS	.43	NS	.78
Checklist	NS	NS	.24	.26	NS	.21	NS	.25
Levels Test	NS	NS	.50	.62	.30	.41	.41	.61
TOEFL	NS	.47	.49	.66	NS	.51	NS	.52
TOEFL Vocab/ Reading	NS	.39	NS	.51	NS	.49	NS	.43

All correlations r<.05

If we examine the data from the perspective of how well the vocabulary tests capture the two kinds of component word knowledge, the large number of nonsignificant correlations indicate no test was consistently successful. The likely reason is that all of the tests focus on only two kinds of word knowledge, written form and meaning, while neglecting other types, like associative and verbal suffix knowledge.

If we accept the assumption that the 'depth of knowledge' of individual words is related to overall vocabulary size, then we can use the association and verbal suffix scores as a benchmark from which to evaluate the vocabulary tests. The results are the most favorable for the Levels Test, as it correlated most consistently with the word knowledge measures and the correlations were relatively strong. The EVST only correlated significantly with 3 out of 8 categories, yet it produced the strongest correlation on the table. The results also confirm the checklist test's weakness (no correlation above .26). Interestingly, the TOEFL correlations were very roughly in line with those of the Levels Test. In contrast, the TOEFL V/R correlations were again lower than the overall TOEFL correlations, further raising suspicions about the vocabulary component. The fact that the TOEFL V/R section, which contains a discrete vocabulary component, correlated less strongly than the overall TOEFL with measures of two different kinds of word knowledge suggests that additional research into the TOEFL vocabulary component is needed.

Two other points warrant a brief mention. The correlations were uniformly more robust on Test 2 (end of school year) than on Test 1 (beginning). As the tests and subjects were identical, this raises the question of whether the improved correlations stemmed from the generally higher language proficiency and larger vocabularies of the students on Test 2. If so, this would suggest that vocabulary size, component word knowledge, and language proficiency become more tightly related as they become more advanced. The second point is that the vocabulary tests correlated more strongly with association knowledge than with verbal suffix knowledge. This might be explained by the fact that association knowledge is closer than suffix knowledge to conceptual meaning, the main type of word knowledge the vocabulary tests measured.

Examining the Various Test Results for Individual Students

It is useful to look at the results for individuals, since important information can often be hidden in group analyses. Table 3 shows the individual vocabulary size estimates from the three vocabulary tests, as well as the TOEFL and TOEFL V/R scores.

Table 3: Individual Vocabulary Size Estimates and TOEFL and TOEFL V/R Scores

	EVST		Checklist		Levels		TOEFL		TOEFL V/R	
	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
1	3514	4520	3708	4400	4444	3756	443	420	41	39
2	4384	4320	3598	3893	4767	4111	430	507	42	47
3	5325	6334	4148	4700	5211	6756	470	533	45	53
4	2562	3644	3654	4366	3700	4744	417	467	39	46
5	2600	2545	4000	2798	3422	3089	403	487	38	48
6	3514	4494	2616	3654	3444	3767	430	460	42	45
7	4588	3749	4148	4257	5656	5289	487	530	49	54
8	1000	3313	2219	3236	3856	3933	407	443	34	43
9	2374	----	1727	----	5167	----	440	---	41	--
10	3554	3615	2981	3598	3744	3367	437	480	41	43

11	2744	3749	3819	3516	4678	5056	473	513	46	49
12	3399	2614	3046	2519	3056	3611	430	440	41	44
13	3534	3759	4100	2923	3800	4167	450	490	39	47
14	3524	5424	3413	3893	5389	5711	453	507	45	50
15	3671	3765	2184	3636	3589	4022	413	467	38	45
16	2420	3660	3200	3516	3056	4200	453	453	38	40
17	3618	3690	2847	3377	3322	3378	437	473	41	49
18	2349	3498	2673	2810	3167	3622	393	430	37	45
19	4518	4624	3774	4475	5033	5244	527	553	50	54
20	2620	3514	4000	4100	2944	3422	417	467	43	44
21	3450	----	2673	3500	3200	3656	403	397	38	37
22	2560	3574	3400	4400	3244	4000	390	403	38	39
23	4364	----	3376	3819	4289	4211	463	467	47	45
24	3510	----	2665	3893	3756	3511	440	467	39	46
25	3514	5465	3114	----	3989	----	427	487	39	45
26	3375	3724	2926	3598	3556	3644	433	440	40	37

Examining the individual results reveals that the different tests produce sometimes quite different vocabulary size estimates. A particularly disturbing trend is that for many subjects, one test showed the subject's vocabulary size had increased, while another test indicated that it had decreased. Even if the weak written checklist test is discounted, the phenomenon remains between the other three tests. Of the 21 Japanese high-school students who took both the T1 and T2 sittings of the Levels, EVST, and TOEFL V/R tests, 7 received scores reflecting conflicting directions of change. Of these, three subjects (Subjects 1,10,12) had EVST and Levels Test scores which were contradictory. In cases where the EVST and Levels Test scores were in agreement, their direction of change was different from the TOEFL V/R test in four cases (2,5,7,26). This variability is disturbing, as the tests should have been at least in agreement regarding the direction of change of vocabulary size.

As for the magnitude of change, it was a bit inconsistent across tests. Although scores from the EVST and Levels Test generally indicated a similar amount of vocabulary size change, some subjects had large increases on the EVST while having much smaller gains on the Levels Test (8,14) and vice versa (15). (The TOEFL V/R is reported in Z-scores, so do not provide a direct estimate of vocabulary size.) This data suggests that while these two tests are unable to provide a rather precise estimate of vocabulary size, they can provide a useful, if somewhat broad, estimate.

It is also interesting to compare the vocabulary test scores with the overall TOEFL proficiency scores (excluding the TOEFL V/R because of score linkage). The TOEFL scores went up in all but two cases (1,21). Of these, the Levels score dropped once and rose once, while one EVST score rose and one was missing. In the 23 cases where the TOEFL rose, in eight cases one or both of the EVST and Levels Test scores dropped. In the three cases where both of these vocabulary scores dropped, the TOEFL score showed a strong improvement of 40 points or more. These incongruencies either indicate weaknesses on the part of the tests, or suggest that the relationship between vocabulary size and language proficiency is not as strong as previous correlation data showed.

Conclusion

Where one might have expected rather high levels of agreement between tests, the analyses showed instead a surprising amount of variability. The scores from the various tests did not correlate together particularly strongly. The EVST and Levels Test usually agreed on the direction of change in vocabulary size, but the magnitude of the change sometimes differed quite substantially. On the other hand, these two tests had quite reasonable correlations with the independent word knowledge measures, showing they are tapping into a learner's 'depth of knowledge' to some extent. The most valid conclusion may be that although the vocabulary tests examined here are our field's current best effort, their imperfections highlight a serious need for further research into vocabulary testing.

The variation indicated by the data in Table 3 suggests that we should view total vocabulary size as something always in flux, where words are forgotten as well as gained. Discovering how to measuring a lexicon's dynamic nature is part of the challenge facing vocabulary testing research. Using types of component word knowledge other than conceptual meaning, such as association knowledge, may well prove a fruitful direction to explore in our quest to meet this challenge.

Note

[1] Length considerations permitted only a very concise report of the study. See Schmitt (n.d.) for the complete account.

References

Cuff, N.B. (1930): "Vocabulary Tests." in *Journal of Educational Psychology 21*, 3: 212-220.

Educational Testing Service. (1987): *Test of English as a Foreign Language*. Princeton, NJ: Educational Testing Service.

Meara, P. (1992): *EFL Vocabulary Tests*. Swansea: Centre for Applied Language Studies, University College, Swansea.

Meara, P. and Buxton, B. (1987): "An alternative to multiple choice vocabulary tests" in *Language Testing 4*, 2: 142-154.

Meara, P. and Jones, G. (1990): *The Eurocentres' 10K Vocabulary Size Test*. Zurich: Eurocentres.

Nation, P. (1983): "Testing and Teaching Vocabulary" in *Guidelines 5*, 1: 12-25.

Nation, I.S.P. (1990): *Teaching and Learning Vocabulary*. New York: Newbury House.

Schmitt, N. (1994): "Vocabulary testing: Questions for test development with six examples of tests of vocabulary size and depth." in *Thai TESOL Bulletin 6*, 2: 9-16.

Schmitt, N. "Verbal Suffix and Word Association Knowledge of Japanese Students." Unpublished MPhil Thesis. University College, Swansea.