Developing an Integrated Diagnostic Test of Vocabulary Size and Depth

Tomoko Ishii and Norbert Schmitt

RELC Journal 2009; 40; 5
DOI: 10.1177/0033688208101452

Article

Developing an Integrated Diagnostic Test of Vocabulary Size and Depth

Tomoko Ishii
Rikkyo University
tmk_ishii@mac.com

Norbert Schmitt
University of Nottingham
Norbert.Schmitt@nottingham.ac.uk

Abstract ■ Following growing interest in vocabulary acquisition, a number of researchers have proposed how learners' vocabulary knowledge can be measured both in terms of how many words they know (vocabulary size) and how well they know those words (depth of knowledge). However, most of the depth measures have addressed only a single depth aspect (often for research purposes), and not many researchers have investigated how size and depth approaches can be combined in a test battery suitable for pedagogical purposes. Furthermore, there is little advice available on how the scores of size and depth measures can be appropriately interpreted. This article describes how one size-depth vocabulary test battery was developed for a specific student group (Japanese university students), and proposes a scoring scheme that combines size and depth scores in a principled way. It also suggests a method of making the resulting scores accessible to the students. It is hoped that the test battery and procedure described can act as a useful guide for teachers and test writers in other countries and contexts who wish to develop a more comprehensive vocabulary assessment approach.

Keywords ■ assessment, depth, size, vocabulary.

Introduction

Vocabulary is now recognized as an essential element of learning a second language, and one which needs to be addressed in a more principled manner than is often the case (Nation 2001). Part of a principled approach to vocabulary teaching involves a better awareness of learners' vocabulary

Vol 40(1) 5-22 | DOI: 10.1177/0033688208101452

http://RELC.sagepub.com

knowledge, and particularly deficiencies in that knowledge. Traditionally, vocabulary knowledge has been conceptualized in terms of *vocabulary size*, that is, the number of words that students know. Size tests are useful in illustrating learners' vocabulary knowledge, especially in conjunction with research that shows how much vocabulary is required for language use. For example, research shows that learners must know 98–99% of words in discourse to understand it well (Hu and Nation 2000), which means they need to know 5000–7000 word families to be conversant in spoken English and 8000–9000 word families to read a range of authentic texts (e.g., novels or newspapers) (Nation 2006; Schmitt 2008). These vocabulary sizes are daunting targets, but are essential for learners wishing to function at a high level in English, and size tests can be utilized to indicate any potential shortcomings.

However, size tests by themselves can tell only part of the story. A high score on a vocabulary size test does not necessarily indicate that the individual words are known very well. Conversely, a student could know the words very well, but only know a few of them.[1] Moreover, many lexical problems are not directly caused by a small vocabulary. Most language teachers would agree that those listed below are among the ones frequently observed:

1. Students know only a limited number of words (Laufer 2000).
2. Students have limited knowledge of secondary meaning senses (Schmitt 1998).
3. Students have limited awareness of the different derivative forms of a word (e.g., *silly*, *silliness*) (Schmitt and Zimmerman 2002).
4. Students use L1 translations when understanding the meaning of L2 words (Jiang 2004).

Although the first problem relates to vocabulary size, the others involve a lack of mastery of words which are only partially (and in many cases only marginally) learned. This lack of *depth of knowledge* can lead to misuse of vocabulary, such as in the following example: 'The food is very nutrients'.

In this case, the student appears to know the word *nutrient,* but does not know its adjective form *nutritious*. A size test which only includes the form *nutrient* would not identify this problem. Rather, detection of a vocabulary deficiency like this requires vocabulary tests which measure *how well* words are known.

Nation (2001: 27) provides a very useful specification of the various types of word knowledge which must be known about vocabulary, includ-

ing written and spoken form, meaning and association connections, grammatical characteristics, collocation, and contextual constraints like register and frequency. A number of researchers have attempted to develop tests which measure these various aspects of vocabulary knowledge. For example, Read (1993) developed the Word Associates Test to measure association knowledge of words, Gyllstad (2007) created collocation tests for Swedish learners of English, and Laufer and Goldstein (2004) developed a computerized test (CATSS) to give an indication of receptive/productive mastery of the form-meaning link of words in the learner's lexicon. However, most of these efforts have focused on a single word knowledge aspect, and few researchers have looked at the interrelationships between the different aspects of word knowledge. Research which has investigated this issue has tended to use rather time-consuming procedures (e.g., Schmitt 1998), which are not practical in any teaching context. As a consequence, instruments which simultaneously look at various aspects of vocabulary knowledge are not yet available.

This is a problem because research shows that having both an adequately-sized lexicon and having rich knowledge about the words in that lexicon are requisites for efficient and fluent vocabulary use (Nation and Gu 2007; Schmitt 2008). This suggests that both of these facets need to be tested in order to obtain a good picture of learners' vocabulary knowledge. Furthermore, using depth measures by themselves is unlikely to be completely satisfactory. If students perform poorly in such tests, it could mean that they are not developing the various types of word knowledge, or it might be due to a small vocabulary, indicating they are not yet at a stage to start learning different types of knowledge. Thus, the scores in depth tests need to be interpreted in light of whether students have a large vocabulary or not, and so vocabulary size and depth should be considered at the same time.

Developing a test battery which measures the complete lexical specification outlined by Nation (2001) is probably impossible, and is certainly impractical in real world classrooms. However, a reasonable compromise is to target the particular aspects which teachers know are problematic in their own teaching context. For example, in our experience, EFL students in Japan typically exhibit the four problems mentioned in the previous section. Based on this, we decided to develop a test battery which addressed the vocabulary size and the three elements of depth related to those problems. The rest of this article will describe the development of the test battery, norms for the target student population, and an integrative

method of interpreting the scores of the battery. We also suggest a method of reporting the results to students in an accessible manner.

The Four Diagnostic Vocabulary Tests

The Vocabulary Size Test

The best-known and most widely-used vocabulary size test is the Vocabulary Levels Test (VLT), originally developed by Paul Nation, and updated and validated by Schmitt, Schmitt and Clapham (2001). It is a receptive matching test and focuses on four frequency levels (2000, 3000, 5000, and 10000), as well as academic vocabulary. It seems to work well overall, but there have been some indications that Japanese learners can have difficulties with the L2 definitions on the test (Aizawa 1998; Kamimoto 2003). Another problem is that updated versions are still based on somewhat outdated word lists.

Considering the possible difficulty involved in reading the English definitions and the use of outdated word list, we developed a new version of the VLT for the Japanese learners targeted by this study. (See Ishii 2005, for a full report of the development process for all four diagnostic tests.) Given that the definitions of the VLT are already written with simple wordings, with relatively high frequency words, it was infeasible to rewrite the definitions to make them significantly easier for the students. Our solution was to give the meanings with L1 translations. 75 items were randomly selected from the lemmatized list of the British National Corpus. This number of items provided acceptable reliability (see below), while at the same time being practical in terms of time efficiency. The test covers up to 6000 lemmas on the list, with five frequency bands: 2000, 3000, 4000, 5000 and 6000. Each band includes five clusters of items, which means 15 words are tested in each band. A sample cluster from the 2000 word level is illustrated below:

Choose the right word to go with each meaning.

1.	leading	
2.	male	a. 薄い
3.	rich	b. 男性の
4.	thin	c. 金持ちの
5.	traditional	
6.	typical	

(kanji characters for: [a] *usui* – thin, [b] *danseino* – male, [c] *kanemochino* – rich)

The Test of Multiple Meaning Senses for Words

English, like many other languages, has a great deal of polysemy. This can be problematic, as learners are often puzzled with words that they think they know, but that do not make sense in the context. For example, Bensoussan and Laufer (1984) found that learners had more trouble guessing the meaning of polysemous words than the meanings of other words. This can cause problems in acquisition, as Schmitt (1998) found that even advanced learners seldom knew all the meaning senses of polysemous words, and that learning them was a slow and patchy process. As such, it made sense to measure this aspect of lexical knowledge in our test battery.

act ( ) ( )
[1] * 行為 [2] 細胞 [3] 幕 [4] 娘 [5] 利点
(*kanji characters for: [1] *koui*—a thing done [2] *saibou*—cell [3] *maku*—a division of a play [4] *musume*—daughter [5] *riten*—advantage)

In this format, two correct meaning senses are provided, and three distractors which are semantically distant from the target word. The test takers are asked to choose two of the five options. They will be awarded a point only if they choose both of the appropriate answers. With this scoring system, the items can be unambiguously marked as correct/incorrect, with no chance for ambiguous 'split' scores gaining partial credit (i.e., one correct option picked as well as one distractor). The 30 items on this test were sampled from the first 2000 lemmas of the BNC lemmatized list.

The Test of Derivative Word Forms

In many English language programs in all parts of the world, the learners are explicitly exposed to a single form of the target words. Learners may learn this particular form because of the explicit attention, but unless they have a great deal of additional exposure to written or spoken English discourse, they may not learn the other word forms of a word's family which were not explicitly taught. For example, learners may learn *silly*, because it is relatively frequent, and so will probably be explicitly taught (3,498 occurrences in the BNC).[2] But they may not be taught the much less frequent *silliness* (112 occurrences), and will have to read or hear a great deal of text in order to learn it incidentally. Indeed, Schmitt and Zimmerman (2002) found that this incomplete knowledge of word family members was the norm, even for relatively advanced learners. They tested the noun, verb, adjective, and adverb members of academic words (e.g., *minimum, minimize, minimal/minimum, minimally*) and found that most learners knew some, but not all, of the forms.

The test format we prepared for this type of knowledge, after piloting different formats, is illustrated below. Students are asked to write one word form under each part of speech.

Target word	Noun	Verb	Adjective
stimulate	*stimulation*	*stimulate*	*stimulating*
educate			

As the regularity in forming adverbs was found problematic when we piloted the test, and as deleting adverbs did not lower the reliability of the test, we decided to use only three parts of speech as shown above.

The 15 words on the test were sampled from the most frequent 2000 lemmas in the BNC. In the cases where two or more possible word forms exist (e.g., adjective: *educational, educated*), students were asked to write only one of the forms they knew. They were awarded full credit if this answer was appropriate.

The Test of Lexical Choice between Near-Synonyms

The overly heavy use of L1 translations can be a problem too. Many students seem to prefer translation as a learning strategy (Liao 2006), and while this can be very useful in the beginning stages of learning a word, a continuing reliance on translation can block the learning of knowledge (e.g., collocation, meaning nuances between near-synonyms) which comes mainly from engaging with the word in L2 contexts (Schmitt 2008). We see many cases in Japan where students get stuck when reading because they persist with the translation they know, and produce semantically strange sentences when writing an essay. Many students do not seem to realize the differences between near-synonyms which may have the same L1 translation, such as *argue*/*discuss*, *damage*/*hurt*, and *suggest*/*offer*, and some research points out the difficulty involved in learning such differences (Jiang 2004). Although near-synonyms may have a similar meaning, they often have different collocational or register properties, which means that they are not interchangeable (McCarthy 1990: 17). This can lead to lexical choice errors, if students are not aware of these different properties. Thus, our last test was developed to assess students' awareness of such near-synonyms.

The format employed for this test is shown below, where the test-takers are supposed to compare two near-synonyms and choose the one which fits better in a context provided.

Choose one word which fits better in the blank:

1) work 2) job 3) I DON'T KNOW

Speaking English is important to find a _____ these days.

This lexical choice test format has the drawback that there is a 50% chance of choosing the correct answer by guessing. However, piloting showed that to ensure all the test-takers looked through all of the options that the test format asks them to compare, it was necessary to limit the item to the two synonyms with no distractor options. When other options were on the test, the examinees were very easily distracted and many of them could not perform well even when they did know the difference between the two target words. We wanted to make sure the examinees carefully considered the near-synonyms, and so we were forced to compromise on the higher probability of the guessing.

The 18 pairs of near-synonyms on this test were also chosen from the first 2000 lemmas in BNC list. For the selection of the items, resources such as EFL textbooks discussing near-synonyms (Rudzka *et al.* 1981), specialized dictionaries for the use of words (Swan 1995; COBUILD 1992), as well as the experience of EFL teachers in Japan, were used. We prepared three sentences for each pair, and the test contained 54 items in total.

Score Interpretation

A Proposal for Score Interpretation

Because there has been very little research into simultaneous testing of vocabulary size and depth, it is not clear how these two interrelate with each other, or how the results of such tests should be interpreted in relation to one another. However, both facets are clearly part of the same underlying lexical competency, so it makes sense to try and interpret the scores in an integrated manner. This is especially true because unless we know what performance we should expect from the learners in the light of their vocabulary size, the scores from depth measures alone are not very informative.

The approach we pursued in our Japanese context was first to establish baseline data for Japanese learners, in order to gauge the typical performance of learners with different vocabulary sizes. Once the baseline norms were established, teachers would be able to compare their students' scores

with what could reasonably be expected based on the baseline norms. If a student scored considerably lower on a particular depth measure, compared to other students with the similar vocabulary size, the teacher would need to draw this student's attention to that aspect of vocabulary knowledge.

Collection of the Baseline Data

The collection of baseline data involved 523 (293 male; 230 female) university students in Japan, aged between 18 and 23. These students were from six universities, and one junior-college in different parts of the country. They were studying a wide variety of majors, including architecture, computer science, economics, English, and medicine. Depending on the time constraints of the course they were taking, some of the students took all of the four tests in one class, but others took the tests on two or more separate occasions. Because of this, not all students took all four tests.

Analysis of the Baseline Data

The overall results of the four tests are shown in Table 1, which suggests that the four tests are working reasonably well. The learners scored about two-thirds correct on all of the tests except that of derivatives, where they only scored 40% correctly. Clearly, the derivation test was the most difficult of the four, which could confirm Schmitt and Zimmerman's (2002) findings of a weakness in this area, or simply be a result of its being the only productive test in the battery (the others are all receptive tests).

Table 1. *Size and Depth Descriptive Statistics*

	Size	*Multiple Meanings*	*Derivatives*	*Lexical Choice*
N of Items	75	30	60	54
N of Subjects	503	498	492	505
Mean Score	49.56 (66%)	19.12 (64%)	22.12 (37%)	35.42 (66%)
SD	16.01	6.62	11.33	8.94
Cronbach Alpha	.96	.89	.94	.88

In order to interpret the four tests together, it is useful to know how the various aspects of vocabulary knowledge relate to each other. Table 2 shows the correlation matrix of the four tests, based on the scores of the 465 students who took all the four tests. As might be expected, the four types of vocabulary knowledge are highly interrelated with each other.

Table 2. *Correlations among the Size and Depth Scores (N = 465)*

	Size	*Multiple Meanings*	*Derivations*	*Lexical Choice*
Size		.82*	.73*	.67*
Multiple Meanings			.69*	.60*
Derivatives				.61*

*Pearson $p < .01$

Table 2 seems to suggest that as vocabulary size grows, all the other three aspects of vocabulary knowledge grow accordingly. However, a closer look at the data reveals that it is not so straightforward, especially when vocabulary size is taken into account. The students were divided into seven different groups according to their scores in the vocabulary size test. The results are shown in Table 3 and visually illustrated in Figure 1.

Table 3. *Mean Scores of the Three Depth Tests for Groups with Different Vocabulary Sizes*

Group	Range in Size Test (Max=75)	Multiple Meanings (Max=30)			Derivatives (Max = 45)			Lexical Choice (Max = 54)		
		N	Mean	SD	N	Mean	SD	N	Mean	SD
A	Over 70	30	27.23	2.10	29	32.17	4.64	28	45.96	3.93
B	61-70	113	25.09	2.78	113	24.62	6.97	115	40.24	6.84
C	51-60	121	21.38	3.13	122	20.14	6.06	124	36.68	6.51
D	41-50	83	18.29	4.25	81	16.77	5.64	82	33.06	7.10
E	31-40	56	13.73	5.34	53	12.72	5.39	58	30.98	6.37
F	21-30	58	11.12	4.72	59	8.61	5.09	60	25.93	6.96
G	Below 20	24	8.67	4.12	19	8.26	5.18	26	19.04	9.16

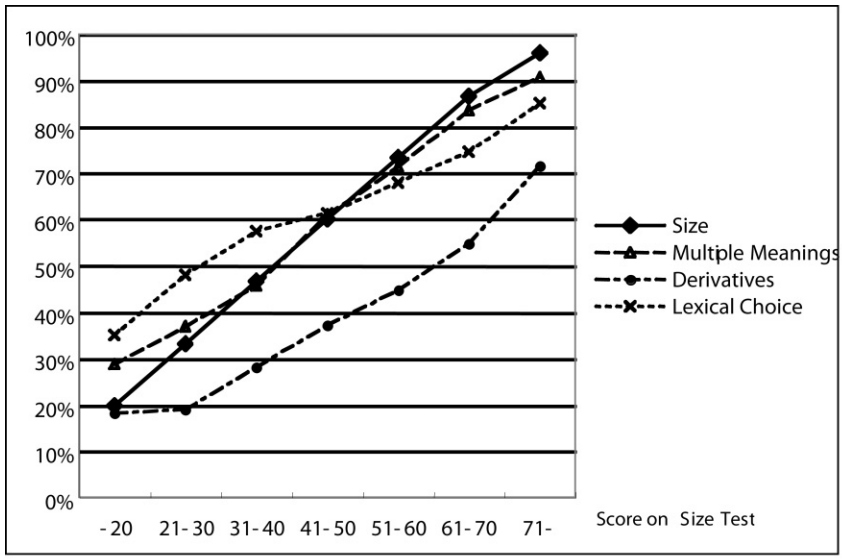

Figure 1. *Relationship between Size and Depth Scores*

Figure 1 shows that although there is a strong positive correlation among the scores of the four tests (Table 2), the interrelationships are not so clear-cut when we break down the data into different vocabulary size groups. Knowledge of multiple meanings does seem to track vocabulary size fairly closely, but knowledge of lexical choice appears relatively higher at lower (<40) vocabulary size levels, and relatively lower at higher size levels (>50). For students scoring 41-50, the scores of the three tests are more or less the same in percentage terms. Derivative knowledge is lower in percentage terms at all vocabulary sizes. Clearly, it is necessary to consider the vocabulary size scores when interpreting the results of the depth tests.

In order to see what performance we can typically expect from students with these different vocabulary sizes, we need to outline the distributions of the scores at each level. Table 4 shows the distributions of the scores of the four tests, taking students scoring between 61-70 on the vocabulary size test as an example. By using the standard deviations, we can make a table to show how the students are distributed in this group (Table 5). We know that about 19.2% of test takers fall between the mean and +0.5SD in a normal distribution, about 15.0% between +0.5SD and + 1SD, and only about 9.2% between +1SD and +1.5SD. The same figures apply to the negative values of SDs. By looking at where a student's scores are placed

in terms of SD, we can detect how typical his or her performance is among the group.

Table 4. *Distribution of Scores for Students Scoring 61-70 on the Vocabulary Size Test*

	Size	*Multiple Meanings*	*Derivations*	*Lexical Choice*
Mean	65.55	25.09	24.62	40.24
	(87.4%)	(83.6%)	(54.7%)	(74.5%)
SD	2.56	2.78	6.97	6.84
Max	70.00	30.00	37.00	52.00
Min	61.00	15.00	1.00	9.00
N	119.00	113.00	113.00	115.00

Table 5. *Standard Deviation Distribution of Students Scoring 61-70 on the Vocabulary Size Test*[a]

	Size	*Multiple Meanings*	*Derivatives*	*Lexical Choice*
+1.5 SD	92.5%	97.5%	78.0%	93.5%
+1 SD	90.8%	92.9%	70.2%	87.2%
+0.5 SD	89.1%	88.3%	62.5%	80.9%
Mean	87.4%	83.6%	54.7%	74.5%
-0.5 SD	85.7%	79.0%	47.0%	68.2%
-1 SD	84.0%	74.4%	39.2%	61.9%
-1.5 SD	82.3%	69.7%	31.5%	55.5%

[a] The figures represent the percentages correct on the various tests

Table 5 will be used as baseline data for interpreting the scores of individual students at this vocabulary size, and similar tables were made for the other vocabulary size groups, e.g. 31-40.

Interpretation of the Vocabulary Test Scores

The tables in the previous section tell us what performance can typically be expected from a student with vocabulary size score of 61-70. Table 6 below shows some example scores from three students from our study.

Table 6. *Example Scores*

	Group Mean		*Student A*		*Student B*		*Student C*	
	Score	%	Score	%	Score	%	Score	%
Size	65.55	87.4%	66	88.0%	63	86.7%	67	89.3%
Multiple Meanings	25.09	83.6%	25	83.3%	23	90.0%	21	70.0%
Derivatives	24.62	54.7%	25	55.6%	21	42.2%	31	68.9%
Lexical Choice	40.24	74.5%	40	74.1%	35	79.6%	42	77.8%

Figure 2 visually illustrates the scores of the same students. These students have similar vocabulary size scores, but have rather different performances on the other three tests. The line for Student A cannot be clearly seen in Figure 2, because this student's performance was very close to the group mean in all four tests. On the other hand, Student B's score in the test of multiple-meanings is higher than the mean score, whereas the score in the test of derivatives is lower. We can see the opposite tendency in Student C. In other words, Student A's performance is what we can typically expect from a student of this vocabulary size, but the other two students' performance is somewhat deviant from the typical performance. Such deviation should draw the teacher's attention, and depending on how severe it is, some remedial steps may need to be taken.[3] In the example above, Student B probably needs to expand their knowledge of derivations, and Student C needs to pay more attention to the different meaning senses of words.

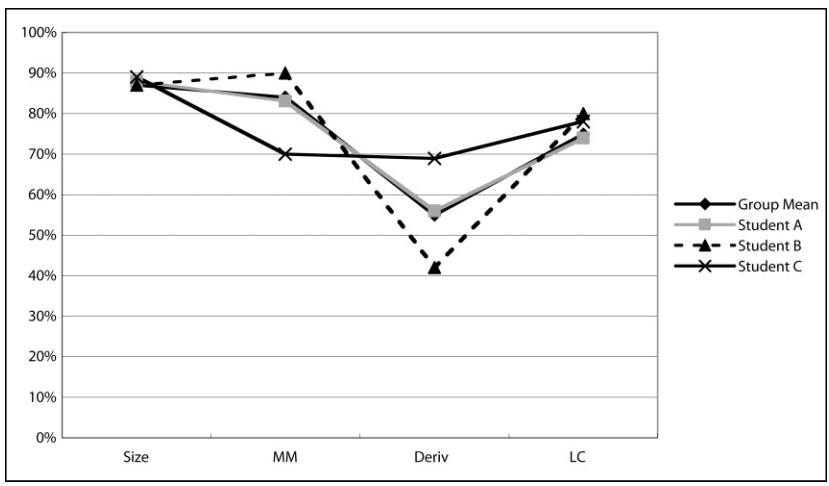

Figure 2. *Comparison of the Example Students' Scores with the Group Means*

As shown above, the scores of the four tests should be interpreted in comparison with the typical performance of the group of students of a similar size of vocabulary. We can plot the performance in the group in a table of distribution of the students. Table 7 below shows the distribution of the scores of the group scoring 61-70 on the size test, with boxes closest to Student C's scores highlighted. In this way, we can easily diagnose the weak aspects of this student.

Table 7. *Distribution of the Scores in 61-70 Group, Highlighted for Student C*

	Size	*MM*	*Deriv*	*LC*
+1.5SD	92.5%	97.5%	78.0%	93.5%
+1SD	90.8%	92.9%	70.2%	87.2%
+0.5SD	89.1%	88.3%	62.5%	80.9%
Mean	87.4%	83.6%	54.7%	74.5%
-0.5SD	85.7%	79.0%	47.0%	68.2%
-1SD	84.0%	74.4%	39.2%	61.9%
-1.5SD	82.3%	69.7%	31.5%	55.5%

Reporting the Scores to Students

A lexical profile like the one described above can clearly be useful for teachers, but the results of the diagnostic tests are meant for students as well. For both end users, the results must be reported in an accessible way which can be easily understood. First, we need to use easy terms, rather than standard deviations, and below is a possible coding (Table 8).

Table 8. *Coding Scheme for Score Interpretation*

Category	*Range*	*Label*
5	Above +1 SD	High
4	1SD – 0.5 SD	Relatively high
3	0.5SD – -0.5 SD	Normal
2	-0.5SD – -1 SD	Relatively low
1	Below –1 SD	Low

Secondly, a format which allows more intuitive interpretation is desirable. One way of doing this is by presenting the information in tables, but it may be more obvious to present the information visually by using a

radar chart to express the balance of the scores, as in Figure 3. We use Student A as an example.

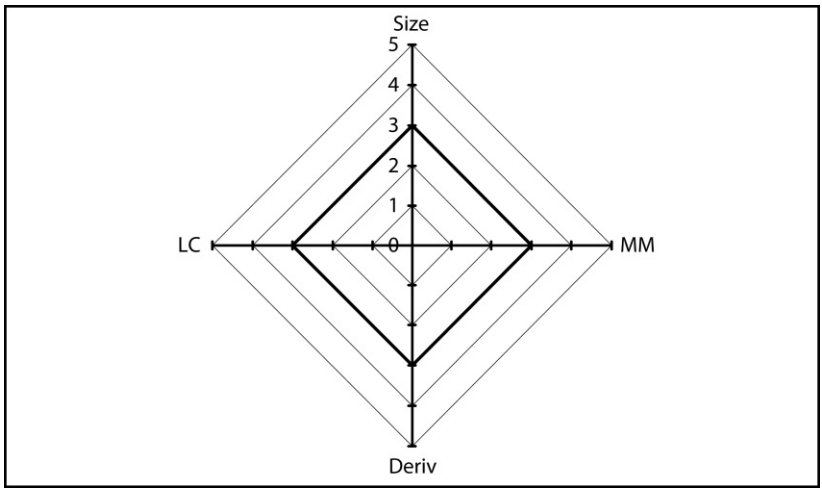

Figure 3. A Radar Chart for the Interpretation of the Test Scores (Student A)

As Student A's performance is close to the means in every test, the chart is perfectly balanced. For students with some deviation, the chart looks quite different. For instance, Figure 4 is the chart for Student B, and Figure 5 for Student C.

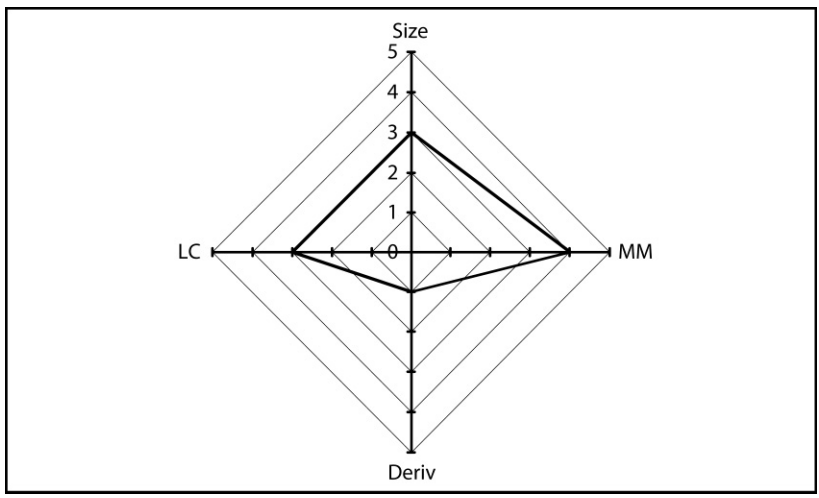

Figure 4. *A Radar Chart for the Interpretation of the Test Scores (Student B)*

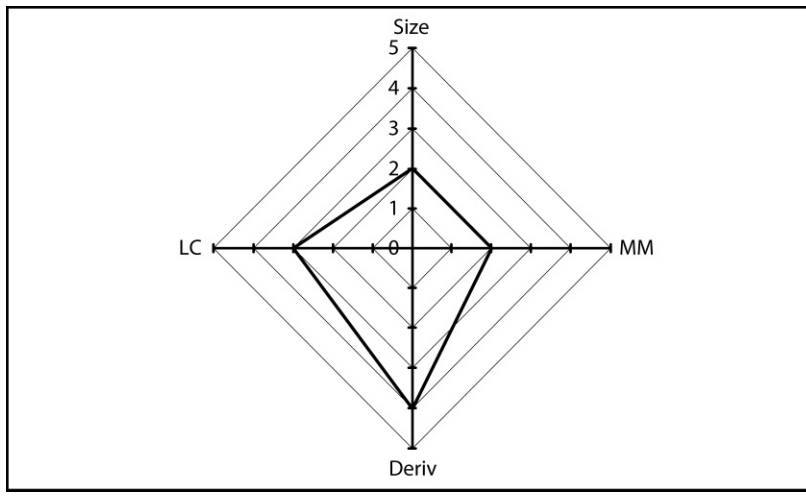

Figure 5. *A Radar Chart for the Interpretation of the Test Scores (Student C)*

These charts are skewed because the scores of the four tests of Students B and C are not balanced. It should be easy for the students to see what their weak areas are. Of course, students will also need to receive an explanation of what their chart means, as well as some advice on steps to take to overcome their weaknesses

Conclusion

Results from research has shown that depth of knowledge is a key component of overall lexical competency, and so should be included in lexical assessment along with vocabulary size. However, although various tests of vocabulary depth have been presented in the field, they have mainly focused on single aspects of word knowledge, and have often not been designed for pedagogical application. There has been little previous work on how to combine size and depth tests in a practical manner, and as a step forward to address this issue, this study proposed a way to interpret the scores from different tests in an integrative manner, using four vocabulary tests as example. Setting vocabulary size as the baseline, we can make tables showing the typical performance of students of a particular vocabulary size, by using the mean scores and the standard deviations. By making a table expressing the distributions of the scores, we can see whether a student's score is higher or lower compared to other students with a simi-

lar vocabulary size. This way, it is possible to diagnose any weak areas of their vocabulary knowledge. The results can then be made accessible to the students by the use of a radar chart. This chart allows us an intuitive interpretation of the four scores in comparison with the typical performance of other students with a similar size of vocabulary.

In sum, we propose that a principled way of diagnosing vocabulary weaknesses is by comparing student performance to the norms of their peers. Vocabulary learning is affected by a wide range of factors, and so it makes sense to compare performance to others who exist in the same environment with its particular possibilities and constraints. Thus this approach calls for baseline data to be collected for each student population. While this certainly requires effort, we believe that the resulting testing procedure can prove very informative for teachers and students alike.

As for our Japan-based test battery, although over 500 students were involved in the study presented in this paper, we would like to expand the baseline data with a larger number and a wider variety of students. Also, test development is an endless challenge, and there is always room for the improvement of the four tests presented in this article. However, we hope that the test battery and procedure presented in this study shows what can be done in combining size and depth tests, and inspires teachers and test writers to pursue a more integrative approach to vocabulary testing.

Notes

1. Research shows that vocabulary size and depth tend to grow in parallel (Vermeer 2001). However, correlations between the two in various studies do not approach 1.00 (e.g., .67-.82 in this study), which indicates that size ≠ depth, and so need to be tested separately.

2. It is also the word family member that is usually the headword in dictionaries.

3. The discussion of remedial teaching is beyond the scope of this article, but see Nation (1990), Nation and Gu (2007), Sökmen (1997), and Thornbury (2002) for useful advice.

REFERENCES

Aizawa, K.
1998 'Developing a Vocabulary Size Test for Japanese EFL Learners', *Annual Review of English Language Education in Japan* 9: 75-85.
Bensoussan, M., and B. Laufer
1984 'Lexical Guessing in Context in EFL Reading Comprehension', *Journal of Research in Reading* 7(1): 15-32.
COBUILD
1992 *English Usage* (Glasgow: Harper Collins).
Gyllstad, H.
2007 'Testing English Collocations' (PhD thesis, Lund University).
Hu, M., and I.S.P. Nation
2000 'Vocabulary Density and Reading Comprehension', *Reading in a Foreign Language* 23(1): 403-30.
Ishii, T.
2005 'Diagnostic Test of Vocabulary Knowledge for Japanese Learners of English' (PhD thesis, University of Nottingham).
Jiang, N.
2004 'Semantic Transfer and Development in Adult L2 Vocabulary Acquisition', in P. Bodaards and B. Laufer (eds.), *Vocabulary in a Second Language* (Amsterdam: John Benjamins Publishing): 101-26.
Kamimoto, T.
2003 'A Comparison of the Vocabulary Levels Test in L1 and L2', *Kumamoto Gakuen University Journal of Language and Literature* 9/10: 217-45.
Laufer, B.
2000 'Task Effect on Instructed Vocabulary Learning: The Hypothesis of "Involvement"' (Selected Papers from AILA '99 Tokyo; Tokyo: Waseda University Press), pp. 47-62.
Laufer, B., and Z. Goldstein
2004 'Testing Vocabulary Knowledge: Size, Strength, and Computer Adaptiveness', *Language Learning* 54(3): 399-436.
Liao, P.
2006 'EFL Learners' Beliefs about and Strategy Use of Translation in English Learning', *RELC Journal* 37(2): 191-215.
McCarthy, M.
1990 *Vocabulary* (Oxford: Oxford University Press).
Nation, I.S.P.
1990 *Teaching and Learning Vocabulary* (New York: Heinle & Heinle).
2001 *Learning Vocabulary in Another Language.* (Cambridge: Cambridge University Press).
2006 'How Large a Vocabulary Is Needed for Reading and Listening?', *Canadian Modern Language Review* 63(1): 59-82.
Nation, P., and P.Y. Gu
2007 *Focus on Vocabulary* (Sydney: National Centre for English Language Teaching and Research).

Read, J.
1993 'The Development of a New Measure of L2 Vocabulary Knowledge', *Language Testing* 10(3): 355-71.
Rudzka, B., J. Channell, Y. Putseys and P. Ostyn
1981 *The Word You Need* (London: Macmillan).
Schmitt, N.
1998 'Tracking the Incremental Acquisition of Second Language Vocabulary', *Language Learning* 48(2): 281-317.
2008 'Instructed Second Language Vocabulary Learning', *Language Teaching Research* 12(3): 329-63.
Schmitt, N., and C.B. Zimmerman
2002 'Derivational Word Forms: What Do Learners Know?', *TESOL Quarterly* 36(2): 145-71.
Schmitt, N., D. Schmitt and C. Clapham
2001 'Developing and Exploring the Behaviour of Two New Versions of the Vocabulary Levels Test', *Language Testing* 18(1): 55-88.
Sökmen, A.J.
1997 'Current Trends in Teaching Second Language Vocabulary', in N. Schmitt and M. McCarthy (eds.), *Vocabulary: Description, Acquisition, and Pedagogy* (Cambridge: Cambridge University Press): 237-57.
Swan, M.
1995 *Practical English Usage* (Oxford: Oxford University Press).
Thornbury, S.
2002 *Teach Vocabulary* (Harlow: Longman).
Vermeer, A.
2001 'Breadth and Depth of Vocabulary in Relation to L1/L2 Acquisition and Frequency of Input', *Applied Psycholinguistics* 22: 217-34.