VOCABULARY TESTING: QUESTIONS FOR TEST DEVELOPMENT WITH SIX EXAMPLES OF TESTS OF VOCABULARY SIZE AND DEPTH

By Norbert Schmitt
Minatogawa Women's College, Japan

Introduction

Although there has been some interest shown in vocabulary testing throughout this century (Sims, 1929; Cronbach, 1943; Dale, 1965; Perkins and Linnville, 1987), the recent surge of attention in vocabulary studies (Meara, 1987; Carter and McCarthy, 1988; Coady, 1993) has given impetus to several fresh testing approaches. Unfortunately, these approaches have not yet filtered down to all classroom teachers, many of whom seem tied to traditional ways of thinking of and testing vocabulary. Although vocabulary achievement tests (tests which measure whether students have learned the words which they were taught in a class or course) remain largely unchanged, improved testing methods have been developed to measure vocabulary size. Perhaps more importantly, work is beginning on an emerging area of vocabulary testing - measuring how well individual words are learned (depth of knowledge), as opposed to the traditional *Yes, the word is known/No, it is not known* dichotomy. This paper aims to help teachers with little or no testing background improve their understanding of vocabulary testing. It will attempt to do this by first proposing a set of principles, in the form of questions, which may prove useful in guiding the writing of better vocabulary tests. Next, several tests of vocabulary size will be examined. Finally, several experimental tests which have potential for measuring learners' depth of knowledge will be discussed. A major theme that will run throughout the paper is that teachers can write better vocabulary tests if they have a clearer understanding of precisely what aspects of word knowledge they wish to test.

Four Questions For Developing A Vocabulary Test

1. WHY DO YOU WANT TO TEST?

This question could be rephrased as "What use will you make of the resulting test scores?" There are several possible purposes for giving a vocabulary test. Perhaps the most common one is to find out if students have learned the words which were taught, or which they were expected to learn (achievement test). Alternatively, a teacher may want to find where their students' vocabularies have gaps, so that specific attention can be given to those areas (diagnostic test). Vocabulary tests can also be used to help place students in the proper class level (placement test). Vocabulary tests which are part of commercial proficiency tests, such as the TOEFL (Educational Testing Service, 1987), attempt to provide a measure of a learner's vocabulary size, which is believed to give an indication of overall language proficiency. Other possibilities include utilizing tests as a means to motivate students to study, to show students their progress in learning new words, and to make selected words more salient by including them on a test. Having a clear idea of which of these purposes the test will be used for can lead to more principled answers to the following questions.

2. WHAT WORDS DO YOU WANT TO TEST?

If the teacher wants to test the students' class achievement, then the words tested should obviously be drawn from the ones covered in class. It is better to avoid standardized tests in this case, because unless an instructor teaches solely from a single book, any general-purpose test is unlikely to be as suitable to a particular classroom and set of students as one the instructor could custom-make (Heaton, 1988). The teacher is in the best position to know her students and which words they should have mastered.

Vocabulary tests used for placement or diagnostic purposes may need to sample from a more general range of words (Heaton, 1988). If the students to be tested all come from the same school, or have been taught from similar syllabi, then it is possible to draw words from those taught in their courses. However, if students come from different schools with different syllabi and language teaching methodologies, as may be the case in a university placement situation, then the words must be more broadly based. In these cases, words are often taken from word frequency lists. These lists were created by counting how frequently various words appeared in a very large collection of written texts (Thorndike and Lorge, 1944; West, 1953; Kucera and Francis, 1967). Since students can generally be expected to know more frequent words best, regardless of their previous schooling, use of these lists allow the principled selection of target words which can be adjusted for students' anticipated language level. The results from tests based on these lists can supply information not only about how many words are known, but also at what frequency level. Tests based on word frequency lists can also be used both within a school system.

Vocabulary tests which are part of proficiency tests need to include the broadest range of words of all. Many universities rely on commercial proficiency tests to control admissions. Therefore, the tests must include a range of words which will provide a fair evaluation of people of different nationalities, native languages, and cultures, as well as proficiency levels. Some of the words on these tests must be uncommon enough to differentiate between higher level test takers.

3. WHAT ASPECTS OF THESE WORDS DO YOU WANT TO TEST?

After the words to be tested have been chosen, the next step is to decide which aspects of those words will be tested. Perhaps the first decision to be made is whether to measure the size of a student's vocabulary (breadth of knowledge) or test how well he knows individual words (depth of knowledge). Until recently, almost all vocabulary tests measured vocabulary size. The vocabulary components of many commercial tests attempt to give an indication of the overall vocabulary size of the testees. In the classroom, vocabulary achievement tests usually try to measure how many words students know from the subset of words they studied. Placement and diagnostic tests have also commonly measured vocabulary size. If teachers are interested in finding out how many words their students know, they will probably decide to test only the conceptual meaning of words, since vocabulary size tests have traditionally measured only that aspect of word knowledge.

However, Nation (1990) has pointed out that a person must know more than just a word's meaning in order to use it fluently. He lists eight kinds of native-speaker word knowledge: knowledge of a word's meaning, spoken form, written form, grammatical patterns (part-of-speech and derivative forms), collocations (other words which naturally occur together with the target word in text), frequency, associations (the meaning relationships of words ie. *diamond - hard, jewelry, weddings*), and stylistic restrictions (such as levels of formality and regional variation). Viewing vocabulary from this perspective, traditional meaning-based know/don't know tests are inadequate for measuring vocabulary knowledge. Depth of knowledge tests are needed which measure some of these components of word knowledge, as well as how fluently they can be put into use. Reflection on the various types of word knowledge can help a teacher decide more precisely which of those aspects she wants to measure and which test formats are the most suitable for that purpose. For example, if she believes that collocational knowledge is important, she would want to use a test format which can capture that kind of knowledge, such as the Multiple True/False test discussed in the

last section of this paper. Also, as the nature of vocabulary acquisition is incremental, tests which consider word knowledge can allow students to demonstrate the components they possess at a given time, even if they are not in full control of every one.

Another important consideration is whether the words will be tested receptively or productively. Although this distinction is more of a continuum than a dichotomy, most test formats fit more easily into one category or other. Examples of predominately receptive test formats are multiple-choice, true/false, and matching, while tests requiring L1 translations, L2 synonyms or definitions, and fill-in-the-blank are examples of productive tests. When should each be used? There are no hard and fast rules, but if a teacher is mainly interested in having his students recognize target words when reading, then a receptive test is suitable. If students are expected to be able to use the target words in their writing, then a productive test may be more appropriate. Also, it might be better to test newer words, to which the students have not yet had much exposure, with receptive tests, since it is generally considered that accurate production requires more control over word knowledge.

The teacher should also consider the mode of the test. Although the vast majority of vocabulary tests are in the written mode, tests in the verbal mode are also possible; dictation and interviews are just two examples. Test mode is related to another factor - whether the test will measure only vocabulary knowledge or whether it will measure how well vocabulary knowledge can be used in conjunction with other language skills, such as reading and writing. This is important because many test formats require the testee to rely heavily on other language skills to answer the item correctly. Let's look at two examples:

1. Write a sentence illustrating the meaning of *gather*.

2. Listen to the tape and write down the word from the story that means the same as *greedy*.

In Example 1, the student may know the meaning of *gather*, but might not be a proficient enough writer to produce a sentence expressing that knowledge. Example 2 shows a task that tests listening ability as well as vocabulary. These kinds of test formats are fine if the teacher wants to measure the control of a word in a language usage context, but are less suitable if the teacher wants a discrete measure of whether the word's conceptual meaning is known or not. This latter case requires isolating the vocabulary knowledge as much as possible from proficiency in other language skills. Of course, this does not mean that vocabulary tests should be devoid of context. The point is that if teachers want to test mainly conceptual meaning, they should try to minimize the difficulty of the reading, writing, speaking, and listening involved in the test items so that limitations in these language skills do not restrict students' ability to demonstrate their vocabulary knowledge. An example of how to achieve this is to always use words of a higher frequency (more common) in the definitions and sentence/discourse context than the target words being tested.

4. HOW WILL YOU ELICIT STUDENTS' KNOWLEDGE OF THESE WORDS?

This question involves decisions about constructing the testing instrument, based on the answers to the preceding questions. The most important decision is what kind (or kinds) of test format will be used. Since different students may have different preferences and different strengths in testing, it may be a good idea to create a test combining several test formats. Heaton (1988) discusses several types of receptive and productive test formats. If the test is to measure depth of knowledge, the test format needs to be carefully selected to ensure it is conducive to measuring the kinds of word knowledge to be tested. (For examples of this, see

the section on Depth of KnowledgeTests.)

The length of the test should also be considered. For any test, the larger the number of test items, the more accurate a picture it will give of students' knowledge. Consequently, situations in which important decisions are made on the basis of test results would normally call for longer and more comprehensive tests. Some test formats, such as checklist and some matching formats allow a larger number of items to be completed within a certain time period. However, the law of diminishing returns has to be considered, as student fatigue sets in on tests requiring a long period of time. It is also important to ensure that the majority of students can complete all of the test items within the given time period. For many purposes, relatively short tests will suffice. For example, tests given for motivational purposes may only need to be 5-10 minutes long.

The best vocabulary test is one in which a student who knows a word is able to answer the test item easily, while a student who does not know the word will find it impossible or very difficult to provide the correct answer. Teachers should ensure that tests have no misleading questions which would trick students who know a word, but on the other hand, tests should not give away any clues which would help students to guess unknown words. For example, Oller (1979) lists the kind of clues that might give away an answer in a multiple-choice test format: the correct choice is either the longest or shortest option, the opposite of the correct choice is given, the alternatives repeatedly refer to the information given in the correct answer, and ridiculous alternatives are included. The following example illustrates these problems.

> A rain forest is a *luxuriant* environment.
>
> a. abundantly and often extravagantly rich and varied
> b. containing little variation
> c. abundant to some extent
> d. containing monkeys and snakes

Even if a student did not know the target word *luxuriant* in this admittedly extreme example, she could probably guess the correct option *a*. It is longer than the other options and has the 'feel' of a dictionary definition, having been taken directly from Webster's Ninth New Collegiate Dictionary (1987). Distractor options *b* and *c* both focus attention on option *a*, while the last option is too silly to consider. Having a colleague look over a new test is a good way of catching such clues that the test-designer is often too 'close' to notice. In fact, it is always a good idea to have someone take the test before it is used in order to uncover problems before it is too late.

While tests should have no obvious clues to help the test-taker guess, it is important to make sure there is enough context in receptive tests to help students understand which meaning of a word is being tested. Productive tests require even more context to narrow the possibilities down to the word the teacher wants. But it is important to remember the point already raised about limitations in other language skills preventing students from exhibiting their full knowledge of words.

Tests Of Vocabulary Size

Since most teachers are probably aware of several kinds of vocabulary achievement tests, the next two sections will give brief introductions to tests teachers are not likely to be familiar with. This section presents three tests which measure vocabulary size, while the next section introduces three experimental tests which attempt to measure the depth of a student's vocabulary knowledge.

A frequently used method of determining the total size of a person's vocabulary in L1 research studies has been **dictionary**

method tests. They involve systematically choosing words from a large dictionary, ie. the fifth word from every tenth page. These words are then fixed on a test. The percentage of correct answers is then multiplied by the number of words in the dictionary to arrive at an estimate of vocabulary size. Unfortunately, this method has many problems, highlighted by widely varying estimates of native-speaker vocabulary size. A serious problem is that dictionaries of different sizes have been used, leading to inconsistent results. Also, the number of test items compared to the total number of possible words (sample rate) is very low. This method cannot really be recommended for determining the total vocabulary size of L2 learners, especially since better methods are available.

One of these methods utilizes the concept that, in general, more frequent words are learned before less frequent words. Instead of using dictionaries which can vary in size as a source for test words, they are taken from frequency count lists. This method entails selecting one or more frequency lists and deciding on the criteria for picking words from the lists. The words from these lists are commonly split into frequency levels at 1,000 word intervals, although smaller groupings are possible. Words are systematically selected from the levels the testees are likely to know, such as the first 2000 most frequent words for beginners. The format is one where words and definitions are matched. The percentage of answers correct in each level's section is multiplied by the total number of words in that level. The scores from all applicable levels tests can be added together to arrive at a total vocabulary score. The obvious advantage of this method is that information is available about how many words learners know at each level. As such, it has even greater applications as a placement or diagnostic test than a test of total vocabulary size. Another major advantage is that these tests are available. The original **Vocabulary Levels Test** appears in Nation (1990), and a revised version with four different forms per level is now being tested for validity and equivalence (Schmitt and Nation, in preparation).

A variation of the same concept features a completely different test format. Checklist tests use the same procedure in selecting the words to be tested, but the learners are only required to 'check' if they know a word or not. This kind of test means that learners can cover many more words than in tests with other item formats, and achieves a much better sampling rate. The obvious problem is that many subjects might overestimate their vocabulary knowledge and check words they really do not know. To compensate for this, nonwords which look like real words but are not, such as *flinder* or *trebron*, are put into the test along with the real words. If some of these nonwords are 'checked' that indicates that the student is overestimating his vocabulary knowledge. A formula compensates for this overestimation to give more accurate scores. The compensation formula works well if the students are careful and mark only a few nonwords, but if they mark very many, then their scores are severely penalized and the test becomes unreliable. (For more on this method, see Meara and Buxton, 1987). There is a book of these checklist tests available, which includes a scoring table, called the **EFL Vocabulary Tests**[1] (Meara, 1992). There is also a commercial computerized version of this test available, the **Eurocentres Vocabulary Size Test**[2] **(EVST)** (Eurocentres, 1990) which requires about nine minutes per student to complete. As with the Vocabulary Levels Test, either of these tests would be particularly suitable as a placement test.

Depth Of Knowledge Tests

Since the area of testing for depth of vocabulary knowledge is so new, there are not yet many depth tests to examine. In fact, in a recent manuscript, Wesche and Paribakht (in preparation) found only one other depth

test to compare with their own. Their experimental test, the **Vocabulary Knowledge Scale (VKS)**, has students rate how well they know a word on the following scale:

I. I don't remember having seen this word before.

II. I have seen this word before, but I don't know what it means.

III. I have seen this word before, and I think it means _____. (synonym or translation)

IV. I know this word. It means _____. (synonym or translation)

V. I can use this word in a sentence: _____. (if you do this section, please also do Section IV.)

(Wesche and Paribakht, in preparation)

This test combines student self-reports, with production to ensure that students do know the words. This kind of test can give a teacher some indication of where along the acquisition continuum a word exists in a student's lexicon. In addition, because it emphasizes what students know, rather than what they don't know, by allowing them to show their partial knowledge of a word, it may be more motivating than other types of tests. But this test has several weaknesses that need to be addressed. One is that we cannot assume that a word is fully learned from just one synonym or sentence. Another is that receptive knowledge is only tested in the first two steps. Also, the number of words that can be covered by the such a test format is rather limited. Most importantly, the best way to score this test is not yet clear.

Another test which attempts to measure how well learners know a word is **The Word Associates Test** being developed by Read (in preparation). This test has the potential to measure associative and collocational word knowledge, in addition to conceptual knowledge. In it, the target word is followed by eight other words, four of which have some relationship with the target word and four which don't. The related words can be synonyms or words similar in meaning (edit - revise), collocates or words which often occur together (edit - film), or words which have some analytical component relationship (electron - tiny). Learners are asked to circle the words which are related.

edit

arithmetic film pole p u b - lishing

revise risk surface text

(Read, 1993)

The scoring system for this test is yet to be worked out, but must eventually take account of the number of correct association words picked, as well as compensate for the number of incorrect distractors circled. Also, since L2 associations are rather unstable (Meara, 1984), this test might be more suitable for more advanced learners.

Cronbach (1943) suggests a test format which aims to provide a more precise measurement of word meaning. His **Multiple True/False Test** asks several true/false questions about the same word. The following examples combine Cronbach's testing idea with some of Nation's (1990) categories of word knowledge. Although this test was created for this paper and has not been validated, it illustrates an approach to be explored which may prove useful in measuring depth of vocabulary knowledge.

Check each acceptable definition or use of the following words.

run
___ to move with quick steps
___ a run in your hair
___ to run in a race
___ a river runs
___ to run down a debt by paying it
___ to run a business
___ a run in a nylon stocking
___ to score a run in football

tap
___ to tap a telephone
___ a gentle knock
___ to embarrass someone
___ a tap on a sink
___ to hit strongly
___ a tap on a car tire
___ to tap one's fingers
___ a tap on a beer keg

This test has the potential to address the polysemous meanings of a word, as well as offering possible collocations for students to consider the correctness of. Items can be written to capture associative relationships, such as those in the Word Associates Test, or stylistic aspects if they are applicable to a word. However, as in the other tests, there are issues to be worked out. The scoring presents problems, although having students answer Y if they are certain of a positive answer, N if they are certain of a negative answer, and ? if they do not know either way has possibilities. It might be difficult to tell when students are guessing and when they actually know the information. Perhaps having more false options would help in this respect. This test also has a weakness similar to multiple-choice tests, in that plausible false options are difficult to write. In spite of these problems, the main reason for presenting this test is to show that existing testing techniques can be creatively adapted to measure depth of vocabulary knowledge.

Conclusion

Teachers will always be interested in vocabulary size and how many words students learn from a course or unit of study. For this reason, tests which measure vocabulary size will remain important. However, there is also likely to be a growing interest in measuring how well those words are learned. We are now only at the beginning stage in the development of depth tests, as indicated by the weaknesses of the above examples. As better depth tests are devised, we are likely to see hybrid vocabulary tests, where size tests are supplemented with depth components to give a broader indication of a learner's lexical capabilities. It is hoped that the example tests briefly examined in this paper will suggest new ways of looking at vocabulary testing to English teachers and that the development questions discussed will give them a principled way of writing their tests in the future.

Notes

1. The EFL Vocabulary Tests are available from: Centre for Applied Language Studies, University College, Swansea SA2 8PP, United Kingdom.

2. The EVST software is available from: Eurocentres Learning Service, Seestrasse 247, CH-8038, Zurich, Switzerland.

References

Carter, R. and McCarthy, M. (1988). *Vocabulary and language learning*. New York: Longman.

Coady, J. (1993). Research on ESL/EFL vocabulary acquisition: Putting it in context. In *Second language reading and vocabulary learning*. T. Huckin, M. Haynes, and J. Coady (eds.) Norwood, NJ: Ablex. 3-23.

Cronbach, L.J. (1943). Measuring knowledge of precise word meaning.

Journal of Educational Research 36, (7), 528-534.

Dale, E. (1965). Vocabulary measurement: techniques and major findings. *Elementary English*, 42, 895-901.

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

Eurocentres Learning Service. (1990). The Eurocentres vocabulary size test. Zurich: Eurocentres.

Heaton, J.B. (1988). *Writing English Language Tests*.. Harlow: Longman.

Kucera, H. and Francis, W.N. (1967). *A computational analysis of present-day American English*. Providence, RI: Brown University Press.

Meara, P. (1992). *EFL vocabulary tests*. Swansea, UK: University College - Centre for Applied Language Studies.

Meara, P. (1987). *Vocabulary in a second language, Volume 2*. London: Centre for Information on Language Teaching and Research.

Meara, P. (1984). The study of lexis in interlanguage. In *Interlanguage*, A. Davies, C. Criper, and A.R.P. Howatt (eds.) Edinburgh University Press. 225-240.

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

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

Oller, J.W. Jr. (1979). *Language Tests at School*.. London: Longman.

Perkins, K. and Linnville, S.E. (1987). A construct definition study of a standardized ESL vocabulary test. *Language Testing*, 4 (2), 125-141.

Read, J. The word associates test: A measure of quality of vocabulary knowledge. Draft Manuscript.

Read, J. The development of a new measure of L2 vocabulary knowledge. Presentation given at Victoria University of Wellington, July 1993.

Schmitt, N. and Nation, P. *Vocabulary levels tests: Versions A, B, C, and D*. In preparation.

Sims, V.M. (1929). The reliability and validity of four types of vocabulary tests. *Journal of Educational Research, 20* (2), 91-96.

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

Webster's Ninth New Collegiate Dictionary. (1987). Springfield, MA: Merriam-Webster.

Wesche, M. and Paribakht, T.S. *Assessing vocabulary knowledge: depth versus breadth*. Draft Manuscript.

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