Second Language Vocabulary Testing: Taking a Broader Perspective

John Read
Victoria University of Wellington

Vocabulary is always considered to be a very important component of second language learning by learners and classroom teachers. As speakers of a second or foreign language, we have all had the experience of not having the words and phrases we needed to express our thoughts and ideas adequately, in the way that we can in our first language. Nevertheless, the study of second language vocabulary learning went somewhat out of fashion in the 1970s and 1980s among applied linguists and language educators, during the time when the communicative approach to language teaching was very influential. This situation has changed to a large extent since the 1990s, with a wealth of books, articles and research reports being published on various aspects of vocabulary acquisition.

A similar pattern can be seen in the development of tests to measure the vocabulary knowledge of language learners. There has been a renewed interest in different types of vocabulary test over the last ten years. However, in some respects vocabulary testing has not kept up with our changing view of the role of vocabulary in communicative performance. One observation I made recently in writing a review of vocabulary research over the last five years (Read, 2004) is that the focus of most published work is still on learners' knowledge of, and ability to use, individual words. This is obviously a core element of vocabulary study but it needs to be reconciled with two important trends in our field: one is the widespread adoption of communicative tasks as a key element in the design of both curricula and tests, under the influence of the communicative approach to language teaching; and the other is the increasing recognition of the extent to which multi-word lexicalised sequences play a role in ordinary language use. This means that we need tests not only to measure knowledge of words in isolation but also the learners' ability to use their vocabulary in a broad sense for various communicative purposes.

In this paper, I want to explore the implications of these two trends for our understanding of the nature of L2 vocabulary assessment and how it should be carried out.

A FRAMEWORK FOR VOCABULARY ASSESSMENT

In order to explain how our conventional thinking about vocabulary testing needs to expand, I developed a framework for my book *Assessing Vocabulary* (Read, 2004). Let me introduce the framework by looking at some familiar types of vocabulary test.

The first test format involves matching words and definitions, as in this example:

1.	environment		
2.	principle	.....	close detailed study of something
3.	response	.....	a set of laws for a country
4.	factor	.....	surrounding area or conditions
5.	analysis		
6.	legislation		

These items are based on the format of Nation's (2001: 416-424) influential Vocabulary Levels Test. The words in the left column are part of a 10 percent sample of words from the Academic Word List (Coxhead, 2000), and in the right column there are three short definitions. The task is to write the number of the word to which each definition belongs. It is deliberately designed as a simple test format, so that a large sample of words from the list can be covered in a reasonably short period of testing time and the intention is that the score should be interpretable as an estimate of how many words in the list the learner has some knowledge of. Thus, each word is presented in isolation and the definitions are kept as short as possible.

A second type of item is illustrated by the following example:

He told me about his dilemma.

a. a serious disease.
b. a difficult choice.
c. a legal problem.
d. a strange dream.

In this case, the target word *dilemma* is presented in a "context" but it is deliberately a limited, neutral one, which does not give any clue as the meaning of the word – or at least none that would allow a learner who did not know the meaning to distinguish among the four options. This can be seen as a less artificial way to present vocabulary to the learners but the primary focus of the assessment is on whether they know the meaning of the word without any supporting information from the sentence in which the word occurs.

Here is a further example:

As a result of the doctors' strike, the c_____ is closed today.

This gap-filling item is designed to test the ability to recall the word *clinic*. This type of item is often referred to as a test of "productive" knowledge, in the sense that the learners must

supply the word rather than simply recognising what it means. As compared to the previous item, the sentence context has more of a role to play in that the test-takers must understand the first part of the sentence – and in particular the word *doctor* – in order to be able to supply the missing word. However, the item is still very much focused on knowledge of the individual target word and does not really involve a communicative task as such.

The three test items I have just presented exhibit three key characteristics of the conventional approach to vocabulary testing:

> *discrete*: it treats vocabulary knowledge as an independent construct, ie, the purpose is to produce a score which can be interpreted as a measure of the learner's vocabulary knowledge, either in an absolute sense (an estimate of total vocabulary size) or with respect to a set of words that have been studied in the course of instruction. It does not set out to assess vocabulary ability in relation to communicative language use.
>
> *specific*: following from the first characteristic, the focus is on particular target words, which may represent a sample from a specified word list or domain of vocabulary use, or they may just be a selection of words that the students have studied recently in class.
>
> *context-independent*: each target word is presented as an isolated language element or, at the most, within a relatively short sentence, which may or may not give any clue as to the meaning of the word. As we saw in the sample items above, the role of the contextual information can be quite variable.

When we move beyond these types of items to think about vocabulary assessment in relation to more task-based approaches to language teaching, we can add a new characteristic to each of the three I have just defined to produce three dimensions of vocabulary assessment, as in Figure 1.

discrete	< ---------- >	**embedded**
A measure of vocabulary knowledge or use as an independent construct		A measure of vocabulary which forms part of the assessment of some other, larger construct

selective	< ---------- >	**comprehensive**
A measure in which specific vocabulary items are the focus of the assessment		A measure which takes account of the whole vocabulary content of the input material (reading/listening tasks) or the test-taker's response (writing/ speaking tasks)
context-independent	< ---------- >	**context-dependent**
A vocabulary measure in which the test-taker can produce the expected response without referring to any context		A vocabulary measure which assesses the test-taker's ability to take account of contextual information in order to produce the expected response

Figure 1. Dimensions of Vocabulary Assessment (Read, 2000, p. 9)

To complement the discrete approach, we can identify vocabulary assessment as *embedded* if it focuses on the learners' vocabulary knowledge or ability as part of their performance of a language use task, involving one or more of the macroskills of listening, speaking, reading and writing. This can be illustrated by the current version (since July 2001) of the Speaking module in the International English Language Testing System (IELTS). The speaking test is conducted by an examiner who rates each candidate's performance according to four criteria: Fluency and Coherence, Lexical Resource, Grammatical Range and Accuracy, and Pronunciation. In my terms, the rating of Lexical Resource is an embedded measure because the examiners are judging how adequate the candidates' vocabulary is for the information and ideas they need to communicate in the test. Thus, the main construct here is speaking ability rather than vocabulary knowledge, and in fact on the IELTS test report form only the overall Speaking score is recorded, not the four individual ratings from which it is calculated. Embedded assessment of vocabulary is also a common component of analytic scoring of writing test tasks.

In the second dimension of the framework, I distinguish between selective and *comprehensive* assessment of vocabulary. Whereas in conventional vocabulary test items the test designer selects the target words that are the focus of the assessment, a comprehensive approach takes account (at least in principle) of all the vocabulary that the test-taker uses in

the performance of a speaking or writing task. Comprehensive assessment can be undertaken in two ways: quantitatively, by calculating one or more lexical statistics, or qualitatively, by means of a rating scale or some other subjective judgement on the examiner's part. Various lexical statistics have been used in studies of the writing or both native speakers and second language learners; in some cases, they have also been applied to the analysis of learner speech. They include lexical error (the number of vocabulary errors); lexical variation (the proportion of *different* words used in a composition), lexical sophistication (the proportion of unusual or low-frequency words); and lexical density (the proportion of content vs. function words). I have discussed the calculation and use of these statistics in some detail in my book (Read, 2000, Chap. 7). The statistics are quite time-consuming to calculate properly and there is only limited evidence of the extent to which they relate to the overall quality of the learners' task performance. This means that they have largely been used in research, rather than for practical assessment purposes.

The alternative form of comprehensive assessment, on the basis of an overall judgement by the examiner, is exemplified by the IELTS speaking test described above. The extent to which examiners can make a valid rating, particularly in a situation where they are required to attend to various other aspects of the candidates' performance at the same time, is an issue that calls for further investigation.

The selective-comprehensive distinction is one that has caused practical difficulty over a long period of time in research to estimate the relative size of learners' receptive and productive vocabulary knowledge (Melka, 1997). A selective approach is very suitable for measuring receptive knowledge of a sample of words through test tasks involving reading or listening, but it is not possible to target the same words in genuine writing or speaking tasks because the essence of productive language use is that the learners make lexical choices, including the avoidance of words that they do not know or are unsure about. Thus, estimates of receptive and productive knowledge end up being not strictly comparable, unless "production" is assessed in a constrained way by using a gap-filling or translation task, which allows the researchers to select the same target words as in the receptive task. I have discussed this issue in some detail in my book (Read, 2000, pp. 154-157).

CONTEXT-DEPENDENT ASSESSMENT OF ACADEMIC VOCABULARY

With regard to the third dimension of the framework, I have already made some preliminary comments about context-dependent assessment in discussing the gap-filling type of item. The main point to be emphasised here is that it is not merely a matter of whether vocabulary items are presented in isolation or in some kind of context for testing purposes. Rather, what is important is whether the test-takers are required to take account of contextual information in responding to the task.

I would now like to explore the notion of context dependency further by reference to some work I have been doing recently to develop tests of learners' knowledge of the Academic Word List. This is a list of 570 word families that occur with high frequency in academic writing across a whole range of subject areas. They are the kind of words which are often referred to as sub-technical vocabulary. The list was compiled by Coxhead (2000) on the basis of a 3.5 million word corpus of various kinds of academic texts which were likely to be encountered by undergraduates in the various faculties of our institution, Victoria University of Wellington. To be included in the list, words had to meet strict criteria of not only frequency but also range across more than half of the subject areas represented in the corpus. Thus, it can be argued that these are important words for international students preparing to undertake degree studies in English-medium tertiary institutions such as ours, and indeed they receive a great deal of attention in the three-month intensive courses of our English Proficiency Programme. The question is how to assess the learners' knowledge of these words and I have been exploring a variety of test formats for this purpose, using the distinction between breadth and depth of knowledge (see Read, forthcoming).

The basic measure of breadth is the matching format I used above as an example of a discrete, selective, context-independent test. Here is another set of items from the test (which is based on one developed for an unpublished MA research paper at our university by Andrea Flavel):

1.	incentive		
2.	display	.....	a person's property and possessions
3.	allocation	.....	encouragement to do something
4.	estate	.....	an amount given for a special purpose
5.	utility		
6.	abstract		

It assesses whether learners have at least some knowledge of a reasonably large sample of words from the list. This is valuable diagnostic information for teachers who want some general indication of their students' familiarity with the list as a whole. However, given the significance of these words for academic reading, it is also necessary to look for more in-depth measures, such as my word associates format (Read, 1993, 1998), as exemplified by the following items:

occur
accident approach exist happen loose roughly

cultural
ethnic justice positive satisfactory society values

consistent

consider constant include pattern remembering same

In this format, the task is to select the three words (or associates) in each set of six which are semantically related to the target AWL word in bold font. This type of item is an efficient way to focus on different elements of word meaning, by probing the learners' depth of understanding of the target word. In addition, research studies by Qian (1999, 2002) with adult ESL learners in Canada and by Noro (2002) with university students in Japan have shown that a word associates test accounts for a significant additional proportion of the variance in reading comprehension test scores beyond what is accounted for by a matching-type breadth test (although it should be noted that in Noro's study the word associates test had this effect only in the case of students who already had a vocabulary size of 3000 words).

However, in terms of my assessment framework, the word associates test – like the matching one – is context-independent by its very nature: it gives no indication of the ability of the students to deal with academic words when they are encountered in a reading task. It is well known that second language learners often have difficulty interpreting polysemous words, even after they have looked them up in a dictionary. For example, in their study of dictionary use by international students preparing to study at a British university, Nesi and Haill (2002) found that more than half of the students failed to find the correct meaning of at least one unfamiliar word in a self-selected reading text, most commonly because they chose the wrong dictionary entry or the wrong meaning of a polysemous word.

Thus I have been exploring the possibility of a discrete and selective but truly context-dependent test of words in the Academic Word List, to complement the matching and word associates tests I have just discussed. The format is not particularly innovative and can be described as a multiple-choice, selective-deletion cloze, or gap-filling task. Figure 2 presents an excerpt from the experimental version of what I am calling a Words in Context test. I composed the text myself, rather than choosing an existing one, in order to control the difficulty of the vocabulary and sentence structure in the context (or perhaps better the "co-text") so that the focus of the assessment could be on the target words from the list. The main issue in writing the test is how to make the items context-dependent. The initial intention was to have, as the options for each item, phrases expressing possible meanings of the target word so that the test-takers would need to process contextual information to be able to choose the correct option. In Figure 2 perhaps the item which best exemplifies this approach is the one for *exploitation*. However, in the test as a whole, it proved to be difficult to maintain this approach consistently. It seems that AWL words are not as polysemous as expected, at least not to the degree that three phrases can be composed representing actual meanings of the word which are plausible in the particular context of the target word.

Using Wind to Produce Electricity

In the 20th century, modern industrial nations have depended mostly on coal, oil and gas to provide plentiful supplies of electricity and other forms of energy to **maintain** their standard of living. However, there are **considerable** problems with each of these fuels and many governments are providing **incentives** for consumers to change to **alternative** forms of energy. One form that seems to have great **potential** at the moment is wind power, which involves the **exploitation** of strong wind currents to produce electricity.

maintain
- improve
- keep
- provide

considerable
- possible
- interesting
- significant

incentives
- amounts of money
- special programmes
- ways to encourage them

alternative
- less harmful
- more expensive
- not so effective

potential
- a large amount of energy
- strength and effectiveness
- the ability to be successful

exploitation
- a use that is unfair
- making good use of it
- starting to use it

Figure 2. An Extract from the Words in Context Test

For the majority of the items, then, I reversed the logic in a sense. As shown in the other sample items in Figure 2, the options tend to represent meanings which are plausible in the context and the test-takers need to have some specific knowledge of the target word to be able to choose the correct answer.

COMPREHENSIVE ASSESSMENT OF THE IELTS SPEAKING TEST

As a further example of comprehensive assessment, I would like to return to the IELTS speaking test. With my colleague Paul Nation, I am currently working on a research project funded by the IELTS Research Program through IELTS Australia to investigate the lexical dimension of learner performance in the test. As previously mentioned, Lexical Resource is one of the four criteria used by the examiners to assess candidates' speaking ability. Research conducted as part of the process of developing the current version of the test indicated that ratings of this criterion were highly correlated with those for two of the other scales: Fluency and Coherence, and Grammatical Range and Accuracy (Taylor, 2001). However, we are interested in seeing whether a more focused examination of vocabulary use by candidates in the test will reveal the lexical features of their speech in this context and also ways in which performance at different proficiency levels can be lexically distinguished.

First, some details about the test. It is conducted with each candidate by a single examiner and consists of three parts:

Part 1: Interview
The candidate answers questions about him/herself and other familiar topic areas

Part 2: Long Turn
After some short preparation, the candidate speaks for 1-2 minutes on a topic specified by the examiner

Part 3: Discussion
The candidate and examiner discuss more abstract issues and concepts related to the Part 2 topic

In order to achieve greater consistency in the way that the test is delivered worldwide, the examiner is required to follow a "frame", which allows for some flexibility but essentially the examiner presents pre-scripted questions and instructions.

In order to carry out the study, we have obtained 100 audio recordings of actual IELTS speaking tests conducted at test centres around the world in 2001-02. The test performances selected all involved one of four topics for Parts 2 and 3, so that the range of topic-specific

vocabulary could be controlled. In addition, they represented performances at three band levels on the nine-band IELTS scale: Band 8, which is a high level of oral proficiency; Band 6, corresponding to a threshold level where the candidate's speaking ability may be considered adequate for study purposes, although limited in some respects; and Band 4, a level which is clearly inadequate for the requirements of academic study.

We are taking two approaches to the analysis of the spoken language produced by the candidates in the test. The first is to treat the test-taker speech as a mini-corpus, with counts of word frequency to calculate the appropriate measures of lexical density, lexical variation and lexical sophistication. This gives us a perspective on the words that candidates commonly use. Obviously, vocabulary use is influenced to some extent by topic and that is why we have restricted our selection of tapes to candidates who were set four particular topics. Secondly, we plan to go beyond the word-based statistical analyses to include a more qualitative study of how the performance of candidates at Bands 8, 6 and 4 can be distinguished in terms of lexical features that include the use of multi-word formulaic expressions (Wray, 2000). This second part of the investigation is just at the planning stage, so let me note briefly that we want to see whether highly proficient candidates are fluent and flexible in their use of a range of formulaic sequences, as compared to those at lower levels of performance, who might be restricted to a limited number of fixed expressions. Alternatively, as Wray (2002: 206-210) argues, less proficient candidates may exhibit the characteristic tendency of adult second language learners in taking an analytic approach to language, so that they compose expressions word by word in situations where native speakers would simply draw on the appropriate formulaic sequence.

To illustrate what the lexical statistics can reveal, let us look at a couple of analyses of our IELTS speaking test corpus. The first one (see Table 1) was generated by *WordSmith Tools* (Smith, 1998), a software package which can produce frequency lists, concordances and collocational analyses for all the words in a text or a whole corpus.

Table 1. Lexical output of IELTS speaking test candidates, by band level

Band Level (no. of candidates)	**Tokens** mean (s.d.)	**Types** mean (s.d.)
Band 8 (n=14)	1170.2 (381.8)	355.5 (81.8)
Band 7 (n=18)	945.8 (293.7)	236.5 (57.5)
Band 6 (n=18)	849.3 (261.4)	248.5 (48.7)
Band 5 (n=18)	684.9 (234.0)	208.4 (58.4)
Band 4 (n=12)	435.2 (133.4)	156.9 (40.2)

In this analysis, the candidates were classified in descending order according to their band score level. The second column in the table shows the mean number of words (or "tokens") produced by students at each level. There is an obvious pattern that higher proficiency candidates say a lot more within the 11-14 minutes of test time than those who are less proficient. The third column shows the number of *different* words (or "types") spoken. Here again, the overall pattern is the same, except at Bands 7 and 6, indicating that candidates who were rated more highly tended to use a wider range of words than those at the lower band levels. However, it is important to note that in both columns the standard deviations are large. This shows that there was a great deal of variation in lexical output among individual candidates at each band score level. Thus, these lexical statistics by themselves would not be reliable indicators of the spoken vocabulary ability of a particular learner.

The second analysis uses different software: the Range program developed by my colleague Paul Nation (Nation and Heatley, 1996) to analyse the vocabulary content of texts according to broad frequency levels. The program classifies words according to whether they are among the 2000 high-frequency words in English, they are in the Academic Word List (as described above), or are not in either of these lists. The third category can be used as a broad measure of "lexical sophistication": the proportion of relatively uncommon, low frequency words used in text. I applied Range to the IELTS speaking test corpus, again dividing the candidates according to their band score, and the results are presented in Table 2.

Table 2. Range analysis of words used by candidates in the IELTS speaking test, by band level and word frequency level

Band score	Total types	Range analysis of types		
		1st 2000 words	Academic words	Other words
Band 8	1516	1041 (69%)	166 (11%)	309 (20%)
Band 7	1475	1050 (71%)	150 (10%)	275 (19%)
Band 6	1214	899 (74%)	125 (10%)	190 (16%)
Band 5	1046	811 (79%)	82 (8%)	153 (15%)
Band 4	673	549 (82%)	46 (7%)	78 (12%)

The table shows the kind of patterns we would expect. Candidates at Bands 4 and 5 use a higher percentage of high-frequency words because presumably they have quite a limited vocabulary knowledge, so that they often do not have the specific words and expressions they need to express themselves more adequately. Conversely, those at Bands 7 and 8 use

proportionately more academic words and words of lower frequency, reflecting their larger vocabulary size and their potential for a much wider range of expression.

These two analyses are obviously just a first step towards a more systematic exploration of ways in which various lexical statistics can help to reveal some of the key features of vocabulary use by learners taking the speaking test. Although a test situation is rather different from a normal conversation, the analysis of our little corpus may yield valuable insights into oral vocabulary use, which has been a neglected area of vocabulary studies until recently.

CONCLUSION

These examples of embedded, context-dependent and comprehensive procedures illustrate new directions for vocabulary assessment. It is necessary to develop them to reflect the shifts which are currently occurring in language teaching and assessment towards task-based approaches and an increasing recognition of the importance of multi-word lexical units in ordinary speech and writing. This is not to say that well-established discrete, selective and context-independent vocabulary tests can or should be dispensed with. Learners, and particularly those at the beginning and intermediate levels, need to give priority to building up a good knowledge of the high-frequency words in the language, and the conventional test formats are efficient ways to assess the progress they are making towards that goal. However, such tests should be complemented by lexical measures which evaluate how effectively the learners can handle vocabulary in use, for both receptive and productive purposes. Thus, we need to continue to explore ways of assessing vocabulary knowledge and ability in the broadest sense as they are manifested in the performance of a range of communicative tasks.

NOTES

This is a substantially revised version of a paper entitled “New directions in second language vocabulary assessment”, which was originally presented at the RELC International Seminar on Teaching and Assessing Language Proficiency in Singapore, November 2003.

The analysis of the IELTS speaking test corpus is based on work supported by IELTS Australia Pty Limited pursuant to an IELTS® Research Grant. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author and do not necessary reflect the views of IELTS Australia Pty Limited, its related bodies corporate or its partners.

REFERENCES

Cambridge ESOL (2003). *IELTS specimen materials 2003*. Cambridge: University of Cambridge ESOL Examinations, on behalf of The British Council and IELTS Australia.

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

Melka, F. J. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 84-120). Cambridge: Cambridge University Press.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, P., & Heatley, A. (1996). *Range* [software program]. Wellington: School of Linguistics and Applied Language Studies, Victoria University of Wellington. Downloadable from www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx

Nesi, H., & Haill, R. (2002). A study of dictionary use by international students at a British university. *International Journal of Lexicography 15*, 277-305.

Noro, T. (2002). The roles of breadth and depth of vocabulary knowledge in reading comprehension in EFL. *Annual Review of English Language Education in Japan 13*, 71-80.

Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge. *Canadian Modern Language Review, 56*, 282-307.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*, 513-536.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10*, 355-371.

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (ed.), *Validation in language assessment* (pp. 41-60). Mahwah, NJ: Lawrence Erlbaum.

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

Read, J. (2004). Research in teaching vocabulary. *Annual Review of Applied Linguistics, 24*, 146-161.

Read, J. (forthcoming). Plumbing the depths: How should the construct of vocabulary knowledge be defined? To appear in P. Bogaards and B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition and testing*. Amsterdam: John Benjamins.

Smith, M. (1998). *WordSmith tools*. Version 3.0. Oxford: Oxford University Press.

Taylor, L. (2001). Revising the IELTS Speaking Test. *Research Notes, 4*, 9-11. [EFL Division, University of Cambridge Local Examinations Syndicate.]

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics, 21*, 463-489.

Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge: Cambridge University Press.