3.4 Vocabulary and testing

John Read
Victoria University of Wellington

If vocabulary knowledge is accepted as a fundamental component of second language proficiency, it is natural to expect that one of the primary goals of language testing will be to assess whether learners know the meanings of the words they need to communicate successfully in the second language. Vocabulary testing does indeed have a lengthy history but, from a contemporary perspective, there are three interrelated issues that need to be addressed in determining the appropriate place for vocabulary in language tests. The first is the role of context in vocabulary testing. While one may generally accept that context is indispensable in normal vocabulary learning and use (see Nagy, 1.4), does this mean that it is always invalid to assess learners' comprehension of particular words presented in isolation from a larger linguistic context? Under the influence of communicative approaches to language teaching and testing, current thinking tends towards the view that we should assess the learners' ability to deal with lexical items as they occur in whole texts and discourse tasks.

This leads to the second issue, which is whether there is still a place for vocabulary tests as such, as distinct from having a lexical focus in integrative tests of listening, speaking, reading or writing skills. Perhaps we need to adopt a broader view of what constitutes a vocabulary test, beyond the dominant notion of a measure of learner knowledge of specific words. This in turn raises the third issue: what the theoretical construct is that underlies any kind of vocabulary testing. If there is more to lexical ability than just word knowledge, what are the other components of this ability and how are they related to each other?

The goal of this chapter, then, is to explore each of these issues in turn.

Objective testing In the US

In his recent book on the development of language testing during this century, Spolsky (1995) traces the origins of second language

vocabulary testing back to the period of the First World War. This was the time when the new science of psychometrics was establishing itself as a dominant force in American education and objective tests were being produced for all the subjects of the school curriculum. Spolsky (1995: 40) attributes the first modern language tests to Daniel Starch, who published tests of Latin, French and German in 1916. These tests assessed vocabulary knowledge by means of a list of foreign words to be matched with their English translations. Early multiple-choice vocabulary items followed a similar pattern, with an L2 word in the stem and four or five L1 words as the options.

Vocabulary, along with grammar and reading comprehension, was the aspect of language that was most commonly included in the new objective tests. There were several reasons for its popularity. Words were seen as meaningful structural units that lent themselves particularly well to objective measurement. In addition, from a practical point of view a multiple-choice vocabulary item involved a minimum amount of item writing, especially if the target word was presented in isolation, and it was relatively easy to find the four or five words or phrases needed to form the options. Multiple-choice vocabulary tests proved to be highly reliable and to correlate very well with tests of reading comprehension as well as psychometric measures of intelligence. Thus, vocabulary tests were valued both for their technical qualities and their apparent validity as indicators of language ability in a broad sense.

Standardized objective tests progressively displaced traditional essay examinations in American educational practice from the 1930s on and, until quite recently, vocabulary items have been a routine component of American language tests. The main issue concerning vocabulary items seems to have been whether the words to be tested could be presented in isolation or should always be in a sentence context. Spolsky (*op. cit.*: 87) cites a study by Stalnaker and Kurath (1935), who found very little difference between a vocabulary test with context and one without, although the test-takers had a slight overall preference for the contextual version.

Contextualization: the case of TOEFL

The issue of contextualization of vocabulary has played an interesting role in the history of what has become the most widely administered English proficiency test in the world today: the Test of English as a Foreign Language (TOEFL). This is the test taken by the hundreds of thousands of foreign students who wish to study in North America, and in some ways it is the classic exemplar of the American objective

tradition in language testing. From its inception in 1964 until 1976, the test included a separate vocabulary section, with two different types of multiple-choice item. The first type – sentence completion – provided a short definition in sentence form in the stem, with a blank to be filled by the correct option:

A _____ is used to eat with.
(A) plow
(B) fork
(C) hammer
(D) needle

The other type, called synonym matching, presented a word or phrase in isolation.

foolish
(A) clever
(B) mild
(C) silly
(D) frank
(Pike 1979: 16)

According to Pike (1979), these item types were criticized because they encouraged students to spend time unproductively learning lists of words and their synonyms. Thus, in a study of alternative formats for TOEFL conducted in the early 1970s, Pike devised a multiple-choice item type called words in context, with the target word appearing in a full sentence in the stem. The intention was to have items with good face validity that would encourage the test-takers to respond to them as they would to a normal reading task. For example:

Nutritionists *categorize* food into seven basic groups.

(A) clarify
(B) grind
(C) classify
(D) channel
(Hale *et al.* 1988: 67)

Pike's research showed that the new vocabulary items not only were very reliable but also correlated highly with the reading comprehension items that were in another section of the experimental test. This led to the intriguing question of whether both vocabulary and reading items needed to be included in the test and, if not, which of the two could be dispensed with. The argument in favour of the vocabulary items was their technical efficiency: they measured student ability with a high

degree of consistency within a short period of testing time. On the other hand, reading is such a crucial skill in university study that it would have seemed very strange to have a test of English for academic purposes that did not require the test-takers to demonstrate their ability to understand written texts. In the end, Pike recommended a compromise solution by which both the words in context vocabulary items and the reading comprehension items were included in a new combined section of the test. Pike's recommendation was accepted and implemented in operational versions of TOEFL from 1976 on.

Nevertheless, criticism of the TOEFL vocabulary items continued. At a conference convened in 1984 by the TOEFL Program to review the extent to which TOEFL could be considered a measure of communicative competence, Bachman observed that the vocabulary items 'would appear to suffer from misguided attempts at contextualization' (1986: 81), because the contextual information in the stem sentence was hardly ever required to answer the item correctly; the most effective response strategy was simply to match the underlined word with the correct option, without spending time to figure out what the whole sentence meant. Subsequently, there were recommendations from the TOEFL Committee of Examiners, an advisory body of scholars from outside the test programme, that lexical knowledge should be assessed in a more integrative manner. Henning (1991) conducted a TOEFL-sponsored study in which he evaluated eight different vocabulary formats incorporating varying degrees of contextualization. Technically, the best format overall was the most contextualized one: the words to be tested were embedded in a reading passage, with a four-option multiple-choice item for each one.

Thus, it is not surprising to find that, in the most recent revision of the structure of TOEFL (implemented in July 1995), there is no longer a separate set of vocabulary items. Lexical assessment has been integrated into the reading comprehension section of the test; that is, a certain proportion of the reading items assess knowledge of particular words in the reading passages, as in the following example:

The word 'capture' in line 8 is closest in meaning to

(A) catch
(B) control
(C) cover
(D) clean

(*TOEFL Sample Test* 1995: 28)

However, if this and the other vocabulary items in the sample test are any indication, Bachman's (1986) observation reported above still

applies: the items can generally be answered correctly without reference to the reading passage in which the words occur.

The TOEFL case highlights a number of points that have wider relevance for vocabulary testing.

- Since well-designed multiple-choice vocabulary items have excellent technical characteristics, they are desirable items to include in a language test if one gives priority to reliability and to purely correlational measures of validity. Increasingly, though, validity is defined much more broadly than just how well a test correlates with other tests. One specific concern of language teachers has been the inclusion of relatively decontextualized items in an important test like TOEFL, because it encourages learners to study lists of isolated words at the expense of a wider range of vocabulary acquisition activities.
- There is a close association between vocabulary knowledge and reading comprehension ability that has long been recognized in the literature (see, e.g. Anderson and Freebody 1981, Nation and Coady 1988). The history of TOEFL points to the difficulty of drawing a dividing line between the two for testing purposes. Spolsky (*op. cit.*: 165) quotes from an unpublished paper written in 1954 by John B. Carroll in which he stated that only test items with a single word stimulus should be classified as vocabulary items; with any longer stimulus, one was dealing with a reading comprehension item. Not many people would accept such a restrictive definition, but in general terms it is true that the more a vocabulary test is contextualized, the more reading comprehension may play a role in test performance.
- As indicated by the recommendations from the TOEFL Committee of Examiners to contextualize vocabulary testing, test items focusing on discrete structural elements – whether they be lexical, grammatical or phonological – have fallen out of favour among language teachers and testers, especially in proficiency testing. Since the 1970s, there has been a decisive shift of opinion in favour of test formats which are integrative and communicative in nature, in keeping with the corresponding trends in language teaching practice. Thus, although TOEFL continues to include vocabulary items, many more recently developed language tests do not.

Vocabulary in language testing

Vocabulary tests

To explore further the changing views towards vocabulary testing, we can survey the numerous handbooks on testing for language teachers

that have appeared in the last thirty years. Given the dominance of standardized objective testing in the US, most of the American books (e.g. Harris, 1969; Clark, 1972; Valette, 1977; Madsen, 1983) have continued to include a substantial section on vocabulary testing, presenting the conventional range of relatively discrete items: not just multiple-choice but also matching, picture labelling, filling a blank in a sentence, and the like.

A clearer trend away from vocabulary testing can be seen in books by British authors. As Spolsky (*op. cit.*) notes, British educationalists have always been much more resistant to the allure of objective testing than their American counterparts. In language testing, as in other fields, the traditions of the subjectively marked examination have been maintained to a large extent. It was not until 1970, for instance, that the University of Cambridge Local Examinations Syndicate (UCLES) first included multiple-choice items in one of their examinations, the Lower Certificate of English (now the First Certificate) (Spolsky *op. cit.*: 213). Nevertheless, the most comprehensive British handbook, Heaton's (1975, 1988) *Writing English Language Tests*, treats vocabulary in a similar manner to the American books, with extensive coverage of objective items – although the second edition gives more emphasis to the desirability of testing words in the context of a whole passage.

Other more recent books from British authors pay much less attention to vocabulary. Harrison (1980) gives just a single text-based multiple-choice format for diagnostic use by teachers in the classroom. Carroll and Hall (1985) and Weir (1990), in their volumes on communicative testing, make little reference to vocabulary beyond arguing for the inadequacy of discrete point testing. On the other hand, Hughes (1989) devotes a half-chapter (shared with grammar) to vocabulary testing towards the end of his handbook. He notes that grammar and vocabulary items are attractive to the designers of large-scale proficiency tests because of their technical qualities and the wide range of content that they can cover. However, he doubts that there is a strong case for a separate vocabulary component in other kinds of language test, except in institutions which emphasize vocabulary teaching: 'For those who believe that systematic teaching of vocabulary is desirable, vocabulary achievement tests are appreciated for their backwash effect' (*op. cit.*: 147). This is scarcely a ringing endorsement. And in a book that has a strong communicative orientation, the sample vocabulary items are surprisingly uncontextualized in nature.

Thus, the predominant impressions to be gained from the recent British books are, first, that the validity of vocabulary testing as such is rather dubious and, secondly, to the extent that vocabulary tests

continue to be administered, there is a dearth of fresh ideas on how to design them, except perhaps for the stronger insistence that the lexical items to be tested should be presented in a whole text rather than a single sentence or in complete isolation.

Integrative lexical measures

However, this does not represent the whole story because the assessment of lexical knowledge and ability is embedded in many current language tests that are not labelled as vocabulary measures. For instance, as illustrated by the reading section of the present version of TOEFL discussed above, reading comprehension tests often include items that focus on understanding of the meaning of particular words and phrases used in the text(s) on which the test is based.

The cloze procedure is an integrative type of language test that can be assumed to draw strongly on the test-takers' lexical knowledge. Scholars such as Bachman (1985), Hale *et al.* (1988), Jonz (1990) and Abraham and Chapelle (1992) have sustained a line of investigation into the question of just what it is that individual cloze items measure. In a careful analysis of eight standard (fixed-ratio deletion) cloze tests, Jonz (*op. cit.*) estimates that about 42 per cent of cloze items require responses that are sensitive to the lexical content of the text. Another 34 per cent of the items involve constraints derived from textual cohesion, which can be taken to have a significant lexical component as well. Presumably in modified cloze tests, in which the words to be deleted are selected individually by the test designer rather than on a fixed-ratio basis, the lexical focus of the assessment can be made even stronger. As noted above, Henning (1991) found very favourable evidence to support the use of a multiple-choice cloze as a contextualized vocabulary measure for TOEFL.

The C-test is a derivative of the cloze procedure, created by selecting several short texts and deleting the second half of every second word in each text, as in the following example:

What is so interesting about work that a whole branch of sociology can be devoted to it? In t____ first pl____, no mat____ how affl____ our soc____ becomes, t____ necessity t____ work wi____ still rem____ the cen____ of o____ existence. Seco____ , the nat____ of wo____ is chan____ so rap____ at t____ present ti____ that ma____ people a____ bewildered b____ it, a____ sociologists bel____ they c____ help th____ to avoid many of the mistakes made in the past.

Although the format was originally devised as an overall test of

language proficiency, its role as a vocabulary measure has also been explored. Chapelle and Abraham (1990) found that C-test scores correlated more highly with a vocabulary test than with reading, writing or listening tests. Singleton and Little (1991) used C-tests in French and German to investigate the nature of the L2 lexicon and, more particularly, to demonstrate that both correct and incorrect responses to C-test items are strongly semantically motivated. In a detailed evaluation of the Singleton and Little's C-test use, Chapelle (1994) found that there were arguments both for and against the validity of the C-test as a measure of L2 vocabulary. Certain items seemed to require application of some aspects of vocabulary ability for successful performance, but it was difficult to separate that out from the influence of other sources.

In communicative tests that require the test-takers to demonstrate their speaking or writing skills, the assessment is normally done subjectively by reference to descriptive rating scales. There may be a single integrated scale or a series of them that focus on various components of the learners' performance. In either case, vocabulary is often one of the components that raters are directed to attend to, if not to rate separately. For example, the influential ESL Composition Profile developed by Jacobs *et al.* (1981) incorporates five scales: content, organization, vocabulary, language use [grammar] and mechanics. The scale for vocabulary is as follows:

20–18	EXCELLENT TO VERY GOOD: sophisticated range • effective word/idiom choice and usage • word form mastery • appropriate register
17–14	GOOD TO AVERAGE: adequate range • occasional errors of word/idiom form, choice, usage *but meaning not obscured*
13–10	FAIR TO POOR: limited range • frequent errors of word/idiom form, choice, usage • *meaning confused or obscured*
9–7	VERY POOR: essentially translation • little knowledge of English vocabulary, idioms, word forms OR not enough to evaluate

Figure 1 Vocabulary scale after Jacobs et al. (1981: 30)

An example for speaking comes from the oral sub-test of the Test of English for Educational Purposes (Weir 1990), for which there were six criteria of assessment: appropriateness, adequacy of vocabulary for purpose, grammatical accuracy, intelligibility, fluency, and relevance and adequacy of content. Here is the vocabulary scale:

Adequacy of vocabulary for purpose

0 Vocabulary inadequate even for the most basic parts of the intended communication.
1 Vocabulary limited to that necessary to express simple elementary needs; inadequacy of vocabulary restricts topics of interaction to the most basic; perhaps frequent lexical inaccuracies and/or excessive repetition.
2 Some misunderstandings may arise through lexical inadequacy or inaccuracy; hesitation and circumlocution are frequent, though there are signs of a developing active vocabulary.
3 Almost no inadequacies or inaccuracies in vocabulary for the task. Only rare circumlocution.

Figure 2 Weir op. cit.: 147

These scales combine potentially quantifiable aspects of the test-takers' performance, such as the number of lexical errors and the range of words used, with more purely qualitative dimensions, like clarity of expression or the appropriateness of word choice. Thus, the raters (especially in speaking tests) are faced with quite a complex task if they are to make a reliable assessment of each test-taker's performance. There are also problems in defining the various steps on the scale and ensuring that there are in some sense equal intervals between them. Very little research has been done to determine the relative merits of this kind of analytic scale as compared with more holistic ratings of writing or speaking performance. Elaborate descriptions are of doubtful value if they have little effect on the way that examiners determine their ratings. But certainly such rating systems represent the opposite end of the continuum of vocabulary assessment from that of uncontextualized multiple-choice items.

Tests in L2 vocabulary research

When we shift the focus from language testing to SLA research into vocabulary teaching or learning, we find several innovations in vocabulary testing. These relate to two areas of interest: estimating vocabulary size (also referred to as *breadth* of vocabulary knowledge) and assessing quality of word knowledge (or *depth* of knowledge).

Estimating vocabulary size

The first area of activity has been on the measurement of vocabulary size, which involves estimating the number of words known by particular groups of language users as well as by individual learners. With native speakers, the objective of studies in this area has been to measure the number of words that they know in some absolute sense (Nation and Waring, 1.1), whereas with second language learners the aim is often more narrowly defined in terms of their knowledge of items in a specified list of relatively high frequency words, such as the General Service List (West, 1953).

Nation and Waring (1.1) have discussed the difficulty of obtaining satisfactory estimates of native speaker vocabulary size, with particular emphasis on the sampling problem. Sampling is less of an issue in making estimates of L2 vocabulary knowledge if a word frequency list is used as the sampling frame. From a testing viewpoint, the question is more one of deciding on the appropriate test format to determine whether each word in the sample is known or not. If a reliable estimate is to be made, the sample of words tested needs to be quite substantial: Nation (1993) calculates that a sample based on *Collins English Dictionary* should ideally contain about 600 items. This requirement places a severe constraint on the kinds of test format that can be used. Typically a simple and relatively decontextualized item type has been chosen for vocabulary size tests, so that the test-takers can respond to the required number of words within a reasonable period of time.

The simplest possible format is the checklist (or yes/no test), which has a lengthy history in L1 research (e.g. Sims, 1929; Tilley, 1936). In its original form, the checklist presents the test-takers with a set of words and requires them to indicate with a tick (✓) whether they know each one. Since the format depends purely on self-report, there is an obvious problem with differing interpretations of what 'knowing a word' means, as well as a lack of any means to check whether the learners are overestimating their vocabulary knowledge. To address this latter shortcoming, Anderson and Freebody (1983) devised a new version of the checklist which contains a certain proportion of plausible non-words that follow the norms of English word formation. Claiming knowledge of some of the non-words is taken as evidence that test-takers are overstating their vocabulary knowledge, so the scores of such learners are adjusted downwards to give a more valid estimate of their knowledge of the real words.

Meara and his colleagues (Meara and Buxton, 1987; Meara and Jones, 1988) developed a computerized checklist test for second language learners of English, one that incorporates non-words and samples

real words from various frequency levels of the Thorndike and Lorge (1944) list. The programme operates on a computer-adaptive principle, presenting words selectively to the test-taker until an adjusted estimate of the individual's vocabulary size can be made, up to a ceiling level of 10,000 words. The test was published as the Eurocentres' Vocabulary Size Test (Meara and Jones, 1990). It was seen as a useful tool for language schools, providing an index of the students' overall knowledge of the language to assist in placing them in the appropriate class.

Although the test initially seemed very promising, Meara (1996c) notes some problems that have emerged from continuing experience with this and other checklist tests. First, they do not work well with low-level learners, who respond unpredictably to the non-words. Secondly, they do not perform satisfactorily as measures of the English language ability of learners whose L1 is French, apparently because of the close relationship between the lexicons of the two languages. The third problem is that certain learners obtain very low scores as a result of their overwillingness to claim knowledge of the non-words. Thus, further work is required to refine the test format and gain a fuller picture of its potential as well as its limitations.

Another well-known size measure is Nation's (1990: 261–72) Vocabulary Levels Test. This is a pen-and-paper test that includes a sample of 36 words for each of five frequency levels from 2,000 to 10,000 words, defined primarily by reference to Thorndike and Lorge (*op. cit.*). The test-takers' task is to match half of the words to short definitions of their meaning, as in this sample set:

1	original	
2	private	___ complete
3	royal	___ first
4	slow	___ not public
5	sorry	
6	total	

The purpose of the test is to give classroom teachers a quick, practical way of profiling their students' vocabulary knowledge at the beginning of a course, in order to provide a basis for planning a vocabulary teaching and learning programme either for the class as a whole or for individual learners within it. As with the Eurocentres test, the words are presented in isolation, and in addition the definitions are expressed as synonyms or short phrases, to minimize the demands of the test task for the learners. In an investigation of the validity of the test, Read (1988) found a substantial degree of implicational scaling across the five frequency levels. In other words, as a general rule the learners knew more of the items at the 2,000 word level than they did at the 3,000

word level and then they knew progressively fewer of the items at the three lower frequency levels. There was also some evidence of the effects of learning when the test was administered a second time at the end of a three-month intensive EAP course.

The measures discussed so far in this section focus on recognition and comprehension of words, or what can be termed breadth of *receptive* vocabulary knowledge. There is also an ongoing series of research studies that investigate the range of vocabulary use in written compositions by L2 learners. Generally speaking, these studies (e.g. Arnaud, 1984, 1992b; Linnarud, 1986; Laufer, 1991) have employed a number of measures of *lexical richness*, including lexical variation (the type-token ratio), lexical density (the percentage of lexical words) and lexical sophistication (the percentage of 'rare' or 'advanced' words). Laufer and Nation (1995) point out various shortcomings of these conventional measures and propose as an alternative the Lexical Frequency Profile (LFP), which reports the percentage of words in the composition that belong to each of four frequency levels. To calculate the LFP, the text of the composition is entered into a computer program, which first classifies the running words into word families (base words plus their inflected and derived forms) and then matches the word families against three frequency lists: the first and second thousand words and the University Word List. The fourth level is composed of word families that are not in any of the three lists. Laufer and Nation (*op. cit.*) present evidence of the reliability and validity of the LFP as a measure of vocabulary size. In another study, Laufer (1994) used the LFP to track the vocabulary development of advanced L2 learners over an academic year. Although some change was evident from the full profile, she found that an increase in productive vocabulary size was more clearly revealed by collapsing the profile into a two-way distinction between the first 2,000 words and the 'beyond 2,000' words.

The need to calculate the LFP by computer means that it is likely to remain a research tool rather than a practical language test, except perhaps where students compose their writing in a computer lab. It shows potential as a measure of the range of productive vocabulary use in the writing of second language learners. However, more work is required to establish how stable an index it is with various types of learners and writing tasks, and also what aspects of vocabulary ability it really represents.

Assessing quality of word knowledge

The other area of recent activity in second language vocabulary testing is quality, or depth, of knowledge of words. Read (1993) has pointed

out that the types of test format used for measuring vocabulary size are inadequate indicators of how well particular words are known, especially in the case of high frequency words that can have a variety of meanings and uses. Read's work has focused on non-technical academic vocabulary, of the kind that is collated in Xue and Nation's University Word List (Nation, 1990: 235–9); these are words that are important for students of English for academic purposes to know well. A second motivation for the development of test procedures to measure quality of knowledge has come from research studies investigating incidental learning of both L1 and L2 vocabulary from reading (Nagy, Herman and Anderson, 1985; Paribakht and Wesche, 1993). These researchers have needed to be able to measure (often small) increases in learner knowledge of words during the course of the study, and clearly a simple yes/no judgement is insufficient for the purpose.

In order to devise tests that assess how much learners know about a word, it is necessary to have some conception of the scope of vocabulary knowledge. There are really two approaches here. One involves analysing the various aspects of meaning and use that characterise 'full' knowledge of a word. In an early article on L1 vocabulary testing, Cronbach (1942) identified what he called five types of behaviour involved in understanding a word: *generalization* (being able to define it); *application* (selecting an appropriate use of it); *breadth of meaning* (recalling its different meanings); *precision of meaning* (applying it correctly to all possible situations); and *availability* (being able to use it productively). He noted that vocabulary tests at that time focused only on the first two: generalization and application. More recently, several writers on second language vocabulary (e.g. Richards, 1976; Nation, 1990: 30–33) have offered lists and frameworks that specify multiple dimensions of word knowledge. Richards' inventory, for instance, includes knowing the relative frequency of a word, its syntactic properties, its underlying form and derivatives, its network of associations with other words, and its connotations. Ideally, then, a vocabulary test might be designed to determine the extent to which each of these aspects of a word was known.

The other approach is a developmental one, identifying levels of knowledge that may be interpreted as stages in the acquisition of the word. Melka (1.5) argues that the traditional dichotomy between receptive and productive vocabulary should be redefined along these lines, as a continuum of degrees of knowledge. Several such scales have been proposed for L1 students, including the one by Dale, who defined four basic stages in knowing a word:

Stage 1: 'I never saw it before.'

Stage 2: 'I've heard of it, but I don't know what it means.'
Stage 3: 'I recognize it in context – it has something to do with . . .'
Stage 4: 'I know it.'
(1965: 898)

Dale also discusses what is in effect a fifth stage: being able to distinguish the word from others that are closely related to it in meaning and/or form, which is akin to Cronbach's (*op. cit.*) 'precision of meaning'. Other scales developed for use with children for whom English is a first language are those of Nagy, Herman and Anderson (*op. cit.*) and Drum (Drum and Konopak, 1987).

Of the two, the developmental approach lends itself better to the construction of tests. Although a scale undoubtedly represents an oversimplification of the multidimensional construct of vocabulary knowledge outlined by Richards (1976) and Nation (*op. cit.*), it is difficult to see how that kind of construct can be operationalized in a practical way, unless knowledge of only a small number of words is to be investigated. The standard procedure for eliciting from learners evidence of how well they know a word is an individual interview, which is obviously a time-consuming procedure and so some scholars have explored the use of written test formats as an alternative.

As part of a study to investigate vocabulary acquisition during a comprehension-based ESL programme at a Canadian university, Paribakht and Wesche (*op. cit.*) developed a written elicitation procedure based on their Vocabulary Knowledge Scale (VKS). The students report their knowledge of each word on a five point scale that is defined for scoring purposes as follows:

1. The word is not familiar at all.
2. The word is familiar but the meaning is not known.
3. A correct synonym or translation is given.
4. The word is used with semantic appropriateness in a sentence.
5. The word is used with semantic appropriateness and grammatical accuracy in a sentence.

The procedure relies on self-report at the first two levels but then requires verifiable evidence for the higher levels of the scale. It also incorporates elements of receptive and productive knowledge, in the sense that the learners are prompted not only to give an explanation of the word but also to compose a sentence containing it. Paribakht and

Wesche found in their study that the VKS was sensitive to gains in the degree of knowledge of the target words during their semester-long ESL course. Subsequently, modified versions of the scale have been used as oral interview procedures by Joe (1995) and Read (1995).

The VKS seems to be a workable instrument, allowing coverage of a reasonable number of words, especially when it is administered as a written procedure. However, various questions can be posed about the validity of the scale. It is by no means obvious that the five levels represent five key stages in the acquisition of a word or that they form an equal-interval scale. For instance, there appears to be quite a large gap between Levels 2 and 3. One can also ask whether supplying synonyms and composing sentences are the most appropriate ways for learners to demonstrate their knowledge of the words. Thus, although the concept of the scale is a good one, it requires considerable refinement to improve its validity.

Another kind of written measure of depth of knowledge, devised by Read (1993), is the *word associates* format, which is based on the concept of word association. In its original form, the test comprised a set of items like the following:

edit			
arithmetic	film	pole	publishing
revise	risk	surface	text

The italicised word is the one being tested. From among the other eight words, the test-takers are required to identify the four that are semantically related to the target word. The four 'associates' are selected to represent various relationships: paradigmatic (synonyms), syntagmatic (collocates) and analytic (representing part of the meaning of the word). In a trial of the format, Read (1993) found that, although the test overall was a good measure of the target vocabulary, there was evidence that the test-takers' willingness to guess played a significant role in their performance on particular items. Specifically, some high-proficiency learners were able to identify associates correctly, without knowing the target word, by looking for semantic links among the associates themselves. In a later study (Read, 1995), a revised version of the format was trialled, together with two concurrent measures: an interview using a modified form of the Vocabulary Knowledge Scale, and a word-definition matching test. Again the test as a whole functioned well, but there was still reason to doubt whether individual item scores really represented how well the corresponding target words were known. Nevertheless, the format is an economical means of assessing the learners' range of knowledge of high-frequency content words.

A broader conception: communicative lexical ability

The discussion of vocabulary testing thus far has been dominated by the concept of vocabulary *knowledge*. Even the ability to use a word in one's own speech or writing is typically referred to as 'productive knowledge'. This is part of the reason for the absence of vocabulary work from the mainstream of contemporary language testing research, because language proficiency is now primarily conceived in terms of the learners' communicative skills and abilities, rather than just their knowledge of the structural elements of the language. This means that, to bring it more in line with current thinking in language testing (and the other disciplines in applied linguistics as well), vocabulary knowledge may need to be reconceptualized within a broader framework of communicative lexical ability, corresponding to Bachman's (1990) model of communicative language ability, which is the most influential construct of second language proficiency at the present time.

One scholar who has explored the nature of such a lexical model is Chapelle (1994, forthcoming). In parallel with the major constituents of Bachman's model, she defines three components of vocabulary ability: '(1) the context of language use; (2) vocabulary knowledge and processes; and (3) the metacognitive strategies required for vocabulary use in context' (1994: 164). From this perspective, most of the work in second language vocabulary testing to date can be seen as focusing on the second component, vocabulary knowledge and processes, which includes vocabulary size, knowledge of word characteristics, organization of the mental lexicon and processes involved in gaining access to the mental lexicon. The result of this focus has been test instruments that assess vocabulary knowledge and use in terms of attributes of the test-takers, without reference to any particular context of use. Thus, words tend to be treated as independent units that can be presented in isolation and whose meaning can be generalized across situations.

This is where the first component of Chapelle's construct, the context of language use, comes in. Under the influence of the concept of communicative competence in applied linguistics, language testers have moved during the last 15 years towards the design of test tasks that incorporate characteristics of what Bachman and Palmer (forthcoming) call 'target language use' (TLU) situations. In other words, test-takers should be given tasks that simulate situations in which they are likely to use the second language outside of the learning environment. It is now a well-established finding from language testing research that test-takers will perform differently, depending on the particular task that is set; this is the so-called 'method effect'. At one level, it means that testing knowledge of words by using multiple-choice items will not yield

exactly the same results as the use of a matching or a yes/no format. At another level, vocabulary test items that lack a discourse context do not allow normal contextual influences to come into play and therefore do not realistically represent the ways in which learners will encounter lexical items in TLU situations. Context has a whole variety of potential effects: on the particular meaning of the word, its connotations, the appropriateness of its use, its interpretability within the linguistic environment, the motivation of the learner to understand it, and so on.

The third component of vocabulary ability is the use of metacognitive strategies, which play a mediating role between the learner's knowledge and processing capacity and the communicative demands of the context. The strategies include circumlocution, paraphrase, language switch, appeal to authority, change of topic and semantic avoidance (Blum-Kulka and Levenston, 1983: 126). According to Chapelle, these strategies:

> . . . call on and manage vocabulary knowledge and processes for all language users, but [they] are particularly important for developing learners who must devise and execute plans for achieving their communicative goals despite limited vocabulary knowledge. (1994: 167)

A definition of communicative lexical ability represents the first step towards a more satisfactory basis for validating vocabulary assessment instruments. Following Messick's (1989) widely accepted framework for test validation, Chapelle (forthcoming) outlines the various kinds of evidence that can be marshalled to support the way that we wish to interpret the results of a particular vocabulary test. These include expert judgements about the content of the test; statistical analysis of how individual items function; verbal report data from test-takers about how they undertake the test task; correlational analyses of different tests guided by theory-based predictions of expected relationships among them; and experimental manipulation of the test-taking situation. Read (1993) explicitly adopted this approach in his investigation of the original version of the word associates test. However, as Chapelle (forthcoming) points out, more work remains to be done to define how the construct validation of a vocabulary test should properly be carried out.

One important component of Messick's framework is consideration of the consequences of using a test. In the case of vocabulary testing, Chapelle (forthcoming) argues that the common use of test formats like yes/no, word association and matching in L2 vocabulary studies has had the effect of narrowing the focus of conceptualization and research, so that researchers have concentrated on knowledge of and access to

individual words, with relatively little attention being paid to the role of contexts and learner strategies. In addition, she sees a danger that decontextualized tests will have a negative washback effect, encouraging learners to concentrate just on studying dictionaries and word lists, to the detriment of their acquisition of a more broadly based lexical ability.

Conclusion

Thus, it seems that the relationship between vocabulary and testing is an uncertain one at present. Although lexical measures continue to be developed and used for a variety of purposes, vocabulary testing has been somewhat on the fringes of both the field of language testing and research on second language acquisition in recent years. Tests of learner knowledge of words that are presented as discrete structural units do not comfortably fit within the dominant testing paradigm, which favours an integrative approach to language test design whereby lexical items are embedded in discourse contexts and the ability of learners is judged according to communicative criteria.

While there is still a role for tests that assess how well learners know words that occur frequently in the language or ones that are useful for the learners' own communicative purposes, these tests need to be located within a broader framework of communicative lexical ability. One priority for such a framework is to specify the role that context should play in vocabulary assessment. As the discussion earlier in this chapter has shown, the role of context has been the most enduring issue in the history of vocabulary testing this century, but context has predominantly been seen as comprising the syntactic structure in which the word is used. Now, a much richer conceptualization of context is required, one that incorporates the multiple linguistic and pragmatic influences that can affect lexical meaning. Another priority is to include the psycholinguistic dimension, involving a specification of the procedures that learners employ to draw on their mental lexicon and the metacognitive strategies that allow them to compensate for deficiencies in their vocabulary knowledge.

In this kind of framework, other lexical measures will be needed in addition to the relatively 'pure' tests of learner knowledge of individual words that currently dominate our thinking about what a vocabulary test is. The future trend in vocabulary testing is likely to be towards the design of integrative test formats that have a strong lexical focus but in which vocabulary ability is one of several factors that contribute to test-taker performance.