The development of a new measure of L2 vocabulary knowledge*

John Read *English Language Institute, Victoria University of Wellington*

With the increasing interest on the part of both researchers and language teaching specialists in the study of vocabulary acquisition by L2 learners, there is a need to have suitable instruments to measure learners' knowledge of words. The purpose of this article is to report on an investigation of a new test format, which was designed to go beyond conventional vocabulary test items – that simply allow for a yes/no judgement on whether a word is known – by providing a practical way of assessing the quality of word knowledge, that is, how *well* particular words are known.

I Design considerations for vocabulary tests

In order to explain the motivation for the development of the new test format, it is useful to review briefly the kinds of format that are currently used in vocabulary testing. An analytic approach is adopted here, involving the identification of four somewhat overlapping dimensions that can be seen to underlie the design of vocabulary test formats.

1 Simple to more complex test formats

The absolute simplest type of item used in vocabulary testing is the yes/no or checklist format (Anderson and Freebody, 1983; Meara and Buxton, 1987), which presents learners with a set of words in isolation and requires them to indicate whether they know each word or not. For example, in the Eurocentres Vocabulary Size Test (Meara and Jones, 1990), which is administered by computer, the test-takers merely press one of two keys to record their response to each word as it appears on the screen. The attraction of the yes/no format is that many words can be covered in a reasonably short period of time. This is particularly desirable in tests to measure vocabulary size, since a large sample size is a significant requirement

*Paper presented at the Language Testing Research Colloquium, Arnhem, The Netherlands, August 1993.

in order to make reliable estimates (Nation, 1993: 35–37). Other commonly used vocabulary item types like multiple-choice or word-definition matching require more reading on the part of the test-taker, especially when the words are contextualized in a sentence or paragraph, with the result that fewer words can be tested within a given period of time. Formats that involve, say, supplying a missing word in a blank or composing a sentence containing the word add a productive writing requirement as well. Obviously, a test based on individual oral interviews is even more time consuming. Thus, test formats vary along this dimension according to the type of response required and the time that needs to be allowed to provide it.

2 *Verifiable responses vs. self-report*

Most vocabulary test formats require test-takers to demonstrate in some verifiable way that they know what the word means. For instance, in a well designed multiple-choice item, only one of the four options is acceptable; in a sentence-writing task, the test-taker must show that he or she can use the word correctly in an appropriate context. However, other formats rely partly or wholly on self-report by the test-taker. In its traditional form, the yes/no test incorporated no check on whether the test-takers were validly reporting their knowledge of the words in the list, which is presumably one reason why researchers came to differing conclusions about the validity of this format (Sims, 1929; Tilley, 1936). Another example of an instrument that relies on self-report is the test of Swedish as a second language developed by von Elek (1985), which presents the test-takers with a range of test items and requires them to indicate on a three-point scale whether they can answer each item correctly or not; they are not asked actually to give the answer.

Although in certain contexts verbal report and self-ratings by test-takers are practical and valid methods of assessing vocabulary knowledge, in most testing situations it is clearly inappropriate to rely on unverified self-report. Thus, modern applications of the yes/no format (see, e.g., Anderson and Freebody, 1983; Meara and Jones, 1990) include a certain proportion of nonwords so that the scores of test-takers who tend to respond 'yes' to items that are not real words can be adjusted downwards accordingly. Similarly, the assessment procedure associated with the Vocabulary Knowledge Scale (VKS) (Paribakht and Wesche, forthcoming) involves a combination of self-report and elicitation of verifiable responses.

3 *Breadth and depth of knowledge*

The terms ‘breadth’ and ‘depth’ can be applied in more than one way to vocabulary knowledge. However, following Anderson and Freebody (1981), breadth of knowledge here refers to the size of a learner’s vocabulary (how many words are known?), whereas depth involves a consideration of the quality of the learner’s vocabulary knowledge (how well are particular words known?). Generally vocabulary tests focus on breadth of knowledge in that they cover as many words as possible within the time allocated and require only a single response in relation to each word tested.

Obviously, though, vocabulary knowledge is very often a matter of degree: even native speakers have only partial knowledge of the meaning of many of the words that they ‘know’. Various researchers have set out to capture this dimension by devising scales to represent different levels of knowledge of words. Paribakht and Wesche’s (forthcoming) Vocabulary Knowledge Scale (VKS) has already been referred to. Other scales of this kind are those proposed by Dale (1965), Nagy, Herman and Anderson (1985), and Drum (Drum and Konopak, 1987). The application of these scales to vocabulary testing requires the elicitation of spoken or written explanations of words from learners, which is a time-consuming procedure that allows for only a limited number of words to be covered.

4 *Testing in context vs. in isolation*

In keeping with the contemporary preference for integrative and communicative test formats, especially in language proficiency testing, it is generally recommended that vocabulary items should be presented in the context of a sentence or a larger discourse unit, since it is important for learners to be able to understand words as they occur in connected written or spoken discourse rather than just as independent items. Nevertheless, the contextualizing of words adds a reading or listening comprehension requirement to the vocabulary test, which in turn reduces the number of items that can be included. For certain purposes, then, particularly in the measurement of vocabulary size (Goulden, Nation and Read, 1990; Meara and Jones, 1990), the desirability of context has to be weighed against the need to cover a large sample of words. This means that presenting words in isolation may be the only practical way of achieving the necessary coverage.

To summarize the discussion, we can say that the design of any vocabulary test represents a compromise among competing considerations such as these. On the one hand, it is normally desirable to cover a reasonable number of words in order to have an adequate sample of the vocabulary to be assessed. This would lead us to prefer a simple and straightforward test format. On the other hand, broad coverage is restricted by a variety of other considerations: the need to verify that test-takers actually know the words that they may claim to know; the desirability of presenting words in some kind of meaningful context; and an interest in establishing how *well* (or how *deeply*) particular words are known. Addressing such concerns may require test formats that are more complex and time consuming to administer.

Thus, the test format presented here represents a new effort to balance the various considerations in the design of a measure of depth of knowledge of words.

II The creation of the format

Initial explorations of depth of vocabulary knowledge by means of an interview procedure (Read, 1989) and a written analogue of the interview yielded interesting results, but demonstrated clearly how few words could be investigated using such methods. This led to the attempt discussed here to develop a vocabulary test format that would involve a simple response task and allow broad coverage of a set of words while, at the same time, probing depth of knowledge of words in some meaningful way.

The starting point for the format was the concept of word association. There is an extensive literature on word associations in L1 (e.g., Deese, 1965; Clark, 1970; Postman and Keppel, 1970), and a significant number of studies have been done with L2 users and learners (Meara, 1980: 234–39; Meara, 1983). One of the basic findings is that native speakers have remarkably stable patterns of word association, which can be taken to reflect the sophisticated lexical and semantic networks that they have developed through their acquisition of the language. On the other hand, second language learners produce associations that are much more diverse and unstable; often their responses are based on purely phonological, rather than semantic, links with the stimulus words. There is some evidence that, as their proficiency increases, the pattern of responses of L2 learners develops towards native speaker norms.

However, it appears that the standard word-association task, in which subjects are asked to provide their own responses to stimulus words, is unsatisfactory as a language test for L2 learners. For one thing, the instability of the learner responses is a major problem from the point of view of the reliability of the test results. Kruse, Pankhurst and Sharwood Smith (1987) experimented with a computer-controlled word association test and found that it was not a good measure of second language proficiency, although the fact that they used only quite small numbers of subjects and words means that any conclusion from their work must be tentative.

The alternative approach adopted here – based on a suggestion by Paul Meara – is to present test-takers with a stimulus word together with a group of other words, some of which are related to the stimulus and some not. The task is to identify the related words, or *associates*; thus, it has come to be known as the *word associates format*. It is assumed that learners with a deeper knowledge of the word will be better able to pick the associates (which should represent different aspects of the meaning of the word) than those whose knowledge is more superficial.

The structure of a word associates item is shown in the following examples:

edit

arithmetic	film	pole	publishing
revise	risk	surface	text

team

alternative	chalk	ear	group
orbit	scientists	sport	together

Preliminary drafting of items showed that there were main types of relationship between stimulus word and associate, as follows:

- *Paradigmatic.* The two words are synonyms or at least similar in meaning, perhaps with one being more general than the other. Examples: edit – revise; team – group.
- *Syntagmatic.* The two words are collocates that often occur together in a sentence. Examples: edit – film; team – scientists.
- *Analytic.* The associate represents one aspect, or component, of the meaning of the stimulus word and is likely to form part of its dictionary definition. Examples: edit – publishing; team – together.

III The development of the test

While the format has potential as a tool in research on the organization of the mental lexicon and the processes by which word knowledge is acquired, the test that is the focus of this study was designed and developed for a more practical purpose: as a measure of vocabulary learning for students of English for academic purposes at university level. Therefore, the sampling frame used for the selection of stimulus words was the University Word List (UWL) (Nation, 1990: 235–39), an 800-word compilation derived from several frequency counts of words in academic English texts. Two systematic samples of 50 words were drawn from the list to be the stimulus words for two forms of the test.

A detailed account of the design of the items and initial try-outs of the test is given in Read (1990). The associates were carefully chosen so that – to the extent that was practicable – they would represent the three types of relationship with the stimulus word outlined above and also cover the main meanings of polysemous words, particularly those meanings that were likely to occur in academic contexts. On the other hand, the distractor words were chosen on the basis that they had no semantic link with the stimulus word. No attempt was made to create 'distraction' by using words that had some orthographic or morphological relation to the stimulus.

When the original set of items was drafted, the associates and distractors were carefully checked to ensure that they were at a higher frequency level – or at least no less frequent – than the stimulus word. The assumption was that each item was testing knowledge of the stimulus word and, from this point of view, it would not be reasonable to present the test-takers with associates and distractors that were more difficult than the stimulus. However, this view was modified to some degree during the process of development, having regard to the purpose of the test as a measure of knowledge of the UWL vocabulary. In order to increase the sampling rate of words from the list, it was decided to take as many of the associates as possible from there. In the majority of cases this meant two UWL associates, but the range was from none to three. In addition, two distractors for each item came from the list. Thus, whereas the stimulus words represented only 6.5% of the list, the addition of the associates and distractors from that source boosted the sampling rate to around 30% for each form of the test.

However, this change in the design principle of the items required some redefinition of what they were measuring. Rather than simply focusing on knowledge of the target word, the items could be seen as assessing the test-takers' ability to identify the lexical network formed by the stimulus plus the four associates. We shall return to this issue of what the items measure in considering the results of the trial, especially the verbal report data.

Preliminary versions of the test were tried out with two groups of students in London. The results presented in this article come from a more formal trial of the test conducted towards the end of the summer English proficiency course at the English Language Institute of Victoria University of Wellington. The majority of the students were preparing for academic study at a New Zealand tertiary institution in the following academic year. Vocabulary study formed a significant component of the learning programme: all of the learners were required either to work with the University Word List itself or to study at least one of the workbooks in the *Advanced English Vocabulary* series (Barnard, 1971–75), which cover a large percentage of the UWL words.

Before being administered to the students, both forms of the test were taken by ten native-speaking teachers on the course. The primary purpose was to confirm the researcher's judgements of which words were associates and which were distractors for each item. In every case where more than one of the teachers did not select an intended associate or chose a distractor instead, the item was reviewed and revised. Fifteen items in form A and 12 in form B were modified in this way, usually by substituting a new associate or distractor. Five of the teachers worked through the revised items and no further problems were found.

Both forms of the test were administered to seven classes, either one after another on the same day or on successive days. An eighth class completed only form B. Before taking the actual test, the students worked through some practice items to familiarize themselves with the format and alert them to the types of association that they should look for. Both on the cover page of the test paper and in the teacher's oral instructions, the students were encouraged to mark responses even if they were doubtful about them. As it said on the cover sheet, 'Try to give as many answers as you can, even if you are not sure whether the answer is correct or not'. In all classes but one, half of the learners took form A first and half form B. Despite the unfamiliarity of the test format, there was no noticeable order effect: in both cases, the scores were substantially higher on form B than form A. One reason for this was that there were a larger number of relatively lower frequency words among the stimulus words in form A.

IV Analyses of the results

Bachman (1990: 243), citing Linn and Messick, emphasizes the need to view test validity as a unitary concept and to gather various kinds of evidence in order to evaluate whether the scores in a test are valid for their intended purpose. Such an approach was followed in this case.

1 IRT analysis

Item response theory (IRT) may be seen primarily as addressing the issue of test reliability. However, Henning (1987: 115) and McNamara (1990) argue that IRT analyses have considerable value as tools in the construct validation of language tests. The responses to each form of the test were analysed separately by means of the partial credit model (Wright and Masters, 1982), a version of the Rasch model that can be applied to test items worth more than one point. For the purpose of this analysis, each of the 50 items was scored out of four, representing the number of associates that were correctly identified. It should be noted here that this method of scoring did not take into account the number of incorrect responses so that, for example, a score of two could be obtained either by making just two correct responses to the item or by making three or four responses, only two of which were correct. The possible significance of this distinction will be taken up in the discussion below. The analysis was carried out using the TITAN program.

There were 103 learners who completed form A and 112 for form B. None of the test-takers obtained a perfect or a zero score on either form of the test, and none of the items was answered either correctly by every test-taker or incorrectly by everyone. Thus, there were no persons or items that were excluded from the analysis for these reasons. The basic statistics for the test are set out in Table 1. The mean scores show that form B was somewhat easier than form A. Both forms of the test achieved a good level of reliability, of at least 0.90.

In order to investigate the relationship between the two forms of

Table 1 Basic statistics from the Rasch analysis of the test

	Form A	Form B
n	103	112
Number of items	50	50
Raw scores		
Maximum score	168	175
Mean	129.57	145.90
Standard deviation	18.27	17.34
Reliability (KR-20)	0.92	0.92
Logit scores		
Mean	1.18	1.56
Standard deviation	0.71	0.77
Reliability (Person separation)	0.91	0.90

the test, the ability estimates from from A were correlated with those from from B for the 101 learners who took both forms. The correlation coefficient (corrected for attenuation) was 0.97, indicating that the two sets of estimates came close to achieving a perfect linear relationship.

A major concern in the application of the Rasch model to test score data is the degree of fit of persons and items to the model. A summary of the infit statistics is presented in Table 2. Misfitting items were identified by reference to the item fit map provided in the output from TITAN. The map plots the infit meansquare values of the items, which should vary around an ideal value of 1.0, with an acceptable range of approximately 0.75 to 1.3 (Tim McNamara, personal communication). By this criterion, three items in form A and one in form B did not fit the model.

Misfitting persons were identified as those with an infit *t* value of greater than +2.0 (Henning, 1987: 123). On this basis, a total of nine test-takers were classified as misfitting on form A and five on form B. Although only two of them were misfitting on both forms of the test, this was an unsatisfactory result, particularly in the case of form A, because the generally acceptable proportion of misfitting persons is 2% (Pollitt and Hutchinson, 1987: 82; McNamara, 1991: 152).

It was not possible to interview the learners identified as misfitting in order to pinpoint reasons for their particular pattern of responding (as suggested by Henning, 1987: 124). However, indirect evidence was sought from an inspection of their test papers and other records. One possible factor was the unfamiliarity of the test format, which might have affected the performance of some test-takers, particularly on the form of the test they took first. The records indicate, though, that this is a plausible explanation in only

Table 2 Summary of fit statistics

	Mean	Standard deviation
Form A		
Infit *t* for items	0.03	0.67
Infit *t* for persons	−0.01	1.24
Form B		
Infit *t* for items	0.05	0.69
Infit *t* for persons	−0.03	1.15
	Form A	Form B
Misfitting items	3	1
Misfitting persons	9	5

about half of the cases at best; the other half of the misfitting test-takers were classified as such on the form of the test that they took *second*.

A review of the test papers of the misfitting persons showed that one common feature of their test-taking behaviour was that they made no response to certain items in the test. Eight of these learners did not respond at all to between three and seven items scattered through the test. Presumably, the reason for not responding was that the learner did not recognize the stimulus word or did not know it well enough to attempt to select associates. This indicates that willingness to guess could have played a significant role in test-taker performance – and further evidence for this is presented below in the discussion of the verbal report data.

Three misfitting persons left a block of items blank towards the end of the test paper, which means that lack of time to complete may have affected their result. The teachers administering the test were told to allow 30 minutes for each form, but they were also encouraged to treat the time limit flexibly and to note how long the learners took. Most completed the test well within 30 minutes. However, one misfitting test-taker noted on his test paper for form B that he had reached only item 39 after 30 minutes. This suggests an overly cautious approach to the test that may be related to the willingness to guess factor noted above.

2 *Analysis of verbal report data*

In addition to the learners who completed the test in class, eight others worked through form A of the test individually in order to provide verbal report data that could give an insight into the process of taking this new kind of test. Bachman (1990: 269–70) points out the usefulness of such qualitative data as a source of evidence that complements conventional quantitative analysis in the construct validation of a language test.

As Cohen (1984: 71–73) has explained, there are various types of verbal report data and considerable debate concerning the most appropriate method of obtaining it. For this study, one learner provided a relatively pure 'think-aloud' protocol by verbalizing her test-taking activity as she sat alone in a room with the tape-recorder. However, it was difficult to find students on the course who were willing to do this and so the other seven subjects worked through the test in the presence of the researcher, who reminded them where necessary to keep verbalizing their thinking as they responded and who elicited what Cohen calls 'immediate retrospection' at certain points during the session to clarify the strategies being employed. It was recognized that this more interventionist approach would have had an influence on the way the subjects took the test, but that disadvantage was outweighed by the more substantial data that was elicited. An audiotape recording was made in each case for later analysis.

The major issue to be considered here is the effect of a willingness to guess responses to stimulus words that were either not known or partially known. Cohen (1984: 74) reported a finding from studies of response behaviour to the cloze test that, whereas less proficient learners were reluctant to guess, more proficient ones were more likely to do so. This was broadly confirmed in this study. Some proficient subjects were indeed resourceful – and quite successful – at identifying associates even though they reported not knowing the stimulus word.

The general strategy followed by all eight subjects was to work through the items in sequence, starting from the first one. In responding to an item headed by a stimulus word they recognized, the subjects scanned the associates and distractors, circling words they could see were related to the stimulus. Typically the subjects were able to identify two or three associates in this manner and then they would look carefully at the remaining words to choose a further response or two. A process of elimination generally ensued: two of the subjects actually put a cross next to words they were confident were not associated with the stimulus. Sometimes the process of elimination left just one low-frequency associate or distractor the learner did not know. At that point the learner had to decide whether to mark the unknown word as a response or not. Whether they did so depended on how sure they were about having eliminated possible alternative responses and their general willingness to guess.

When they encountered an unknown stimulus word, the learners' strategy appeared to vary according to their proficiency level. Three lower-proficiency subjects were more inclined to skip such items, even when the researcher suggested they try to guess what the correct responses were. The items in the test were roughly sorted in decreasing order of frequency and familiarity, so that the last ten items in particular had lower-frequency stimulus words. Thus, the three less proficient subjects were particularly reluctant to guess when confronted with a whole page of these difficult items at the end of the test. On the other hand, learners with higher proficiency were more willing to guess. As one of them put it: 'I don't want to leave it without any answer. I will guess [. . .] because guess has some chance but not guess – nothing!'

Perhaps the most significant finding from the verbal report data was that those who were willing to guess were often quite successful at selecting the correct responses – and in fact they were by no means guessing blindly. The primary reason was that the associates not only were linked individually to the stimulus word but often also had semantic connections among themselves. This meant that test-takers with a good vocabulary knowledge could identify probable associates without knowing the stimulus word, especially when the word had a single basic meaning and context of use. For example, an explicit application of a guessing strategy based on the associates was used by learner no. 4 in item 41 of form A:

41	*denominator*			
	common	develop	divide	eloquent
	fraction	mathematics	species	western

Researcher (R): Do you know this word?

Learner (L): I don't know this word . . . but I'm trying to look at all the words that . . . what are the similar in meaning because I know that four words can go together . . .

[Pause]

I'm not sure that my idea will be right . . .

R: You've circled 'fraction' . . . 'divide' . . . 'species' . . .

L: Or 'develop' . . . So do you want my reason?

R: Yeah.

L: Because I don't know this word before and I tried to looking for some words that have similar in meaning. I think it's a guess – 'fraction', 'divide', 'species' – all things are cut in smaller – yeah – and then it develop bigger amount . . . but I don't know . . .

Similarly, Learner no. 7 guessed three of the associates in item 29 correctly, while emphasizing that he did not know the stimulus word *diffuse*:

29	*diffuse*			
	circulate	government	holiday	light
	optional	scatter	tolerate	vague

L: I don't know what but maybe . . . 'circulate' . . . 'scatter' . . . 'tolerate' . . . 'light' – this I don't know, it's a guess only.

R: So how did you guess? Before you move on, did you have any reason for choosing those words?

L: Reasons for this? I guess this – I read on the words here and I think something like 'circulate' and 'scatter' is – look like the movement – something like that – and from this I decided to guess but I am not sure . . .

This kind of verbal report makes it clear that we should be cautious about interpreting the successful identification of two or even three associates in an item as necessarily evidence that the test-taker knows the stimulus word. It seems that, frequently, the ability to select correct responses is based on a combination of good vocabulary knowledge and a certain resourcefulness in seeking possible associates, as well as the confidence to make guesses.

3 Correlational evidence

In addition to the word associates test, another measure of academic vocabulary knowledge was available for these learners: the vocabulary section of the ELI Proficiency Test (EPT). This measure was a 100-item test of words from the University Word List, using a word-definition matching format, for example:

1	physical	—	a	belonging to the country
2	fertile	—	b	enough for the purpose
3	tropical	—	c	able to produce young
4	adequate	—	d	very important
5	rural	—	e	very large
			f	related to the body
			g	in the hot parts of the world

The EPT battery, including the vocabulary test, was administered in the last week of the course, approximately two weeks after the learners completed the word associates test. Thus the EPT vocabulary test could be used as some kind of criterion measure to estimate the concurrent validity of the word associates test. As with most measures employed for this purpose, it was the best one at hand rather than an ideal criterion measure. IRT ability estimates were not available for the EPT test and so the raw scores out of 100 were used in the calculations. The correlations of the EPT vocabulary scores and the TITAN ability estimates from the word associates test (corrected for attenuation) are as follows:

Form A
$n = 96$
$r = 0.67$ (observed)
$r = 0.76$ (corrected)

Form B
$n = 106$
$r = 0.73$ (observed)
$r = 0.81$ (corrected)

The results indicate that there is a very substantial relationship between the two tests, with a somewhat higher correlation for form

B than for form A. The correlations may be depressed a little by a ceiling effect in the case of the EPT vocabulary test, in which the mean score was 87.9 out of 100 and about a quarter of the test-takers scored at least 95. This analysis, then, can be interpreted as some evidence for the concurrent validity of the word associates test.

V Discussion

The evidence from the various analyses shows that this word associates test is reliable and that it seems to provide an efficient measure of knowledge of academic vocabulary, as represented by the words in the University Word List.

The test results are affected to some extent by the willingness, or otherwise, of test-takers to guess what the associates are, especially when they do not recognize the stimulus word or are not sure what it means. It seems that some proficient test-takers are disadvantaged — and are identified as such by the IRT analysis – by giving no response to certain items. Conversely, those who are willing to guess and supply a full set of responses may enhance their performance on the test. The fact that in numerous items it is possible to identify associates simply by recognizing semantic relationships among them, without reference to the stimulus word, means that one needs to be cautious about interpreting correct responses as evidence of how well the stimulus word is known. Thus it is not clear to what extent this format has achieved the goal of measuring depth of knowledge of particular words. It remains to be seen whether it is possible to write word associates items that minimize the likelihood that test-takers can successfully guess the associates without knowing the meaning of the stimulus word.

One characteristic of the test developed for this study is that the items are heterogeneous in structure in a variety of ways. The stimulus words include nouns, verbs and adjectives, and each word class enters into somewhat different relations with other words. Some words have essentially a single meaning and limited range of use, while others are polysemous and broad in range. This affects the range of possible associates that can be found. In addition, there is little consistency from one item to another in the relationships (syntagmatic, paradigmatic, analytic) between the associates and the stimulus word. This heterogeneity, of course, reflects the actual diversity of words in the language and is appropriate in a test that was designed with a practical application in mind: as a measure of the quality of vocabulary learning in a university EAP course.

Therefore, it was important from this point of view to draw a representative sample of words from the University World List.

However, from a research point of view, the test is rather complex to analyse. The item analysis data provided by the TITAN program shows great variability in the patterns of responses to individual items. This is reflected in the step thresholds, which vary greatly from one item to another. In addition, the analysis does not distinguish between a lack of response to an item and one or more incorrect responses to it. In order to evaluate the potential of the word associates format as a research tool more fully, it will be necessary to develop tests that focus on more homogeneous subsets of vocabulary items so that greater consistency can be achieved in the semantic relationships among the words and in the pattern of responses elicited. Another limitation of the study is the relatively small number of test-takers involved in the trial. Some more conclusive results should be obtained from the analysis of a rather larger dataset.

Acknowledgements

I am grateful to Paul Meara for his role in the devising of the test format and for his support, especially during the early stages of test development. I also wish to thank Tim McNamara for his assistance in carrying out the TITAN analysis and helping to interpret the results.

VI References

Anderson, R.C. and **Freebody, P.** 1981: Vocabulary knowledge. In Guthrie, J.T., editor, *Comprehension and teaching: research reviews.* Newark, DE: International Reading Association.

—— 1983: Reading comprehension and the assessment and acquisition of word knowledge. *Advances in reading/language research. Volume 2.* Greenwich, CT: JAI Press.

Bachman, L.F. 1990: *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Barnard, H. 1971–75: *Advanced English vocabulary. Workbooks 1 – 3B.* Rowley, MA: Newbury House.

Clark, H.H. 1970: Word associations and linguistic theory. In Lyons, J., editor, *New horizons in linguistics.* Harmondsworth: Penguin.

Cohen, A.D. 1984: On taking language tests: what the students report. *Language Testing* 1, 70–81.

Dale, E. 1965: Vocabulary measurement: techniques and major findings. *Elementary English* 42, 895–901, 948.

Deese, J. 1965: *The structure of associations in language and thought.* Baltimore: Johns Hopkins University Press.

Drum, P.A. and **Konopak, B.C.** 1987: Learning word meanings from written context. In McKeown, M.G. and Curtis, M.E., editors, *The nature of vocabulary acquisition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Goulden, R., Nation, P. and **Read, J.** 1990: How large can a receptive vocabulary be? *Applied Linguistics* 11, 341–63.

Henning, G. 1987: *A guide to language testing: development, evaluation, research.* Cambridge, MA: Newbury House.

Kruse, H., Pankhurst, J. and **Sharwood Smith, M.** 1987: A multiple word association probe in second language acquisition research. *Studies in Second Language Acquisition* 9, 141–54.

McNamara, T.F. 1990: Item response theory and the validation of an ESP test for health professionals. *Language Testing* 7, 52–76.

—— 1991: Test dimensionality: IRT analysis of an ESP listening test. *Language Testing* 8, 139–59.

Meara, P. 1980: Vocabulary acquisition: a neglected aspect of language learning. *Language Teaching and Linguistics: Abstracts* 13, 221–46.

—— 1983: Word associations in a foreign language. *Nottingham Linguistics Circular* 11, 29–38.

Meara, P. and **Buxton, B.** 1987: An alternative to multiple choice vocabulary tests. *Language Testing* 4, 142–54.

Meara, P. and **Jones, G.** 1990: *Eurocentres Vocabulary Size Tests 10KA.* Zurich: Eurocentres Learning Service.

Nagy, W.E., Herman, P.A. and **Anderson, R.C.** 1985: Learning words from context. *Reading Research Quarterly* 20, 233–53.

Nation, I.S.P. 1990: *Teaching and learning vocabulary.* New York: Newbury House.

Nation, P. 1993: Using dictionaries to estimate vocabulary size: essential, but rarely followed procedures. *Language Testing* 10, 27–40.

Paribakht, T.S. and **Wesche, M.B.** forthcoming: The relationship between reading comprehension and second language development in a comprehension-based ESL program. *TESL Canada Journal.*

Pollitt, A. and **Hutchinson, C.** 1987: Calibrated graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing* 4, 72–92.

Postman, L. and **Keppel, G.** 1970: *Norms of word associations.* New York: Academic Press.

Read, J. 1989: *Towards a deeper assessment of vocabulary knowledge. ERIC Document Reproduction Service* ED 654 321. Washington, DC: ERIC Clearinghouse on Languages and Linguistics.

—— 1990: The word associates test: a measure of the quality of vocabulary knowledge. Unpublished working paper.

Sims, V.M. 1929: The reliability and validity of four types of vocabulary test. *Journal of Educational Research* 20, 91–96.

Tilley, H.C. 1936: A technique for determining the relative difficulty of word meanings among elementary school children. *Journal of Experimental Education* 5, 61–64.

von Elek, T. 1985: A test of Swedish as a second language: an experiment in self-assessment. In Lee, Y.P., Fok, A.C.Y.Y., Lord, R. and Low, G., editors, *New directions in language testing.* Oxford: Pergamon Press.

Wright, B.D. and **Masters, G.N.** 1982: *Rating scale analysis.* Chicago, IL: Mesa Press.