Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test

Norbert Schmitt *University of Nottingham,*
Diane Schmitt *Nottingham Trent University*
and **Caroline Clapham** *University of Lancaster*

The Vocabulary Levels Test has been widely used in language assessment and vocabulary research despite never having been properly validated. This article reports on a study which uses a range of analysis techniques to present validity evidence, and to explore the equivalence of two revised and expanded versions of the Vocabulary Levels Test.

I Introduction

Vocabulary is an essential building block of language and, as such, it makes sense to be able to measure learners' knowledge of it. This is equally true whether we are interested in pedagogical assessment in classrooms or in language acquisition research. Given this, one might expect there to be an accepted vocabulary test available for these uses. Unfortunately, this is not the case. The closest thing the field has to such a vocabulary test is the Vocabulary Levels Test (Nation, 1983, 1990). Different versions have been employed in both assessment and research, but despite this widespread use this test has never been properly validated. This article aims to begin to address this shortcoming by describing an initial validation of two revised and expanded versions.

The Vocabulary Levels Test is designed to give an estimate of vocabulary size for second language (L2) learners of general or academic English. The rationale for the test stems from research which has shown that vocabulary size is directly related to the ability to use English in various ways. For example, knowledge of the most frequent 2000 words in English provides the bulk of the lexical resources required for basic everyday oral communication (Schonell *et al*., 1956). The next 1000 words provide additional material for spoken

Address for correspondence: Norbert Schmitt, Department of English Studies, University of Nottingham, Nottingham, NG7 2RD, UK; email: norbert.schmitt@nottingham.ac.uk

Language Testing 2001 18 (1) 55–88 0265-5322(01)LT197OA

discourse but, additionally, knowledge of around 3000 words is the threshold which should allow learners to begin to read authentic texts. Most research indicates that knowledge of the most frequent 5000 words should provide enough vocabulary to enable learners to read authentic texts. Of course many words will still be unknown, but this level of knowledge should allow learners to infer the meaning of many of the novel words from context, and to understand most of the communicative content of the text. L2 learners with a knowledge of the most frequent 10 000 words in English can be considered to have a wide vocabulary, and Hazenberg and Hulstijn (1996) found that a vocabulary of this magnitude may be required to cope with the challenges of university study in an L2. For L2 learners of English who wish to engage in an English-medium academic environment, knowledge of the sub-technical vocabulary that occurs across a range of academic disciplines (academic vocabulary) is also necessary. (For an overview of vocabulary size research, see Nation and Waring, 1997.)

The Vocabulary Levels Test provides an estimate of vocabulary size at each of the above four frequency levels and also provides an estimate of the size of the examinees' academic vocabulary. This information can be utilized by teachers and administrators in a pedagogical context to inform decisions concerning whether an examinee is likely to have the lexical resources necessary to cope with certain language tasks, such as reading authentic materials. The information can also be used to identify possible lexical deficiencies which might need addressing. Similarly, results from the Vocabulary Levels Test can be used in research studies where an estimate of lexical size at the relevant frequency levels is considered informative (e.g., Cobb, 1997; Schmitt and Meara, 1997; Laufer and Paribakht, 1998). (An extended discussion of the implications of vocabulary size for vocabulary pedagogy is beyond the scope of this article; for details, see Coady and Huckin, 1997; Schmitt and McCarthy, 1997; Read, 2000; Schmitt, 2000; and Nation, 2001.)

II History of the Vocabulary Levels Test

The Vocabulary Levels Test ('Levels Test' for short) was originally designed by Paul Nation as a diagnostic vocabulary test for use by teachers. It first appeared in 1983 and was later republished in his 1990 book. Read (1988) did some initial validation work on the test, finding it to be reliable and finding that subject scores on the different frequency levels tended to fall into an implicational scale (i.e., knowing lower-frequency words tended to imply knowing higher-frequency ones). However, this work was not followed up by any further studies. Despite this, the test began to be used internationally as

Nation's book became increasingly used as a key vocabulary reference source. In 1993, while visiting the Victoria University of Wellington, the lead author of this article revised the Levels Test in Nation's book (Version A) and wrote three additional versions (Versions B, C and D). However, at that time he was unable to run a validation study on them. Since then, these four versions, as well as the original version from Nation's book, have been used in various institutions as an assessment tool (for examples, see Beglar and Hunt, 1999). They have also been used in a number of vocabulary research studies (e.g., Cobb, 1997; Schmitt and Meara, 1997; Laufer and Paribakht, 1998). Recently, Laufer and Nation (1999) created a productive Levels format, based on Versions A–D. The result is that the Levels Test has become widely used in vocabulary research and as a vocabulary test in situations where English for general or academic purposes is taught to speakers of other languages.

Recently, there have been some preliminary moves to explore the validity of the test. Beglar and Hunt (1999) looked at and revised the (unattributed) 2000 and University Word List sections of Versions A–D (see below for a description of the test), and found that they were essentially measuring a single construct. Beglar and Hunt also reported that scores on their Levels sections correlated with TOEFL scores, and items within sections were strongly related to one another. (We come back to these issues below.) In addition, Kudo (personal communication) is attempting to validate a translated version of the test for the Japanese context. However, given that the Levels Test is being used globally in both assessment and research, a more substantial study of some of the test's characteristics is required. The present study is based on the responses of learners of general or academic English from a number of different nationalities at a number of different test sites.

Messick (1989) suggests that a demonstration of the validity of a test should include both logical argumentation and empirical evidence based on quantitative and qualitative data. The authors therefore administered versions of the Levels Test to 801 learners of English and explored the results via item analysis, profile analysis, factor analysis and an examination of these tests' reliability and equivalence. They also investigated the concurrent validity of the tests by correlating the results with the results of an interview (see the section below on the interview with examinees). In addition to these quantitative procedures, more qualitative procedures were also employed: a number of examinees were interviewed to discover what they thought of the tests. They were also asked retrospectively to report the steps they went through in answering the items. Taken together, we believe the results give at least initial evidence that the Levels Test provides

accurate estimates of the vocabulary size of students at the targeted frequency levels.

III Design aspects of the Levels Test

The Levels Test derives its name from the fact that separate sections measure learners' knowledge of words from a number of distinct frequency levels. In this way, it can provide a profile of a learner's vocabulary, rather than just a single-figure estimate of overall vocabulary size. As mentioned above, the levels addressed are the 2000, 3000, 5000 and 10 000 frequency levels. In addition, there is a section for academic vocabulary. (For a fuller description of the format, see Nation, 1990.)

The frequency counts used were ones commonly available in 1993: Thorndike and Lorge (1944), Kučera and Francis (1967) and the General Service List (GSL) (West, 1953). Words were taken in a stratified sampling from the Thorndike and Lorge list, with reference to frequency data from Kučera and Francis and the GSL. The only exception to this is the 2000 section, where words from the 1000 level and the 2000 level were sampled at a 1:2 ratio. (The first thousand words of the GSL are usually those with a frequency higher than 332 occurrences per 5 million words, plus months, days of the week, numbers, titles (Mr, Mrs, Miss, Ms, Mister), and frequent greetings (Hello, Hi etc).) The words in the Academic section were sampled from the University Word List (Xue and Nation, 1984). (Because the University Word List was not yet available when Nation wrote the original Levels Test, the Academic section of the original test was sampled from Campion and Elley, 1971.)

Reflecting the distribution of these word classes in English, the words from the stratified sample tended to fall into a 3 (noun) : 2 (verb) : 1 (adjective) ratio. This ratio was maintained in the test, with each section containing three noun clusters, two verb clusters and one adjective cluster. The following illustrates the format of a noun cluster:

> You must choose the right word to go with each meaning. Write the number of that word next to its meaning.
>
> 1 concrete
> 2 era _____ circular shape
> 3 fiber _____ top of a mountain
> 4 hip _____ a long period of time
> 5 loop
> 6 summit
>
> [*Note*: the test is written with American English spellings, but test users are free to change these if they wish.]

Each cluster was written with the following considerations in mind:

1) The options in this format are words instead of definitions.
2) The definitions are kept short, so that there is a minimum of reading, allowing for more items to be taken within a given period of time.
3) Words are learned incrementally, and tests should aim to tap into partial lexical knowledge (Nagy *et al.*, 1985). The Levels Test was designed to do this. The option words in each cluster are chosen so that they have very different meanings. Thus, even if learners have only a minimal impression of a target word's meaning, they should be able to make the correct match.
4) The clusters are designed to minimize aids to guessing. The target words are in alphabetical order, and the definitions are in order of length. In addition, the target words to be defined were selected randomly.
5) The words used in the definitions are always more frequent than the target words. The 2000 level words are defined with 1000 level words and, wherever possible, the target words at other levels are defined with words from the GSL (essentially the 2000 level) (for more details, see Nation, 1990: 264). This is obviously important as it is necessary to ensure that the ability to demonstrate knowledge of the target words is not compromised by a lack of knowledge of the defining words.
6) The word counts from which the target words were sampled typically give base forms. However, derived forms are sometimes the most frequent members of a word family. Therefore, the frequency of the members of each target word family was checked, and the most frequent one attached to the test. In the case of derivatives, affixes up to and including Level 5 of Bauer and Nation's (1993) hierarchy were allowed.
7) As much as possible, target words in each cluster begin with different letters and do not have similar orthographic forms. Likewise, similarities between the target words and words in their respective definitions were avoided whenever possible.

IV Issues in vocabulary testing

Before reporting the study, it is first necessary to address several vocabulary testing issues. In particular, we feel that some of the methods commonly used for test validation need to be carefully scrutinized before they are used with vocabulary tests.

An accepted way of exploring validity is to examine correlations of test scores with several other measures, each having varying

degrees of similarity and difference with regard to the test being validated. In this way one can look for both convergent and discriminant patterns of relationship. However, the measures providing convergent evidence each need to address the same construct. To provide this evidence, vocabulary tests are often correlated with proficiency tests, particularly the TOEFL. However, since the TOEFL is more a measure of general proficiency than vocabulary, we do not feel that it addresses the construct of vocabulary knowledge sufficiently well to be a satisfactory criterion. It would be more informative to compare the test with other accepted tests of vocabulary size. However, as mentioned before, the Levels Test is the closest thing we have to an accepted measure (although see the Eurocentres IOK Vocabulary Size Test by Meara and Jones, 1990), and it is the one which gives frequency profile information instead of a single figure for overall vocabulary size. We, therefore, needed to develop a suitable criterion to explore convergent concurrent validity. Although time-consuming, a relatively dependable way of probing vocabulary knowledge is through personal interview, so we designed a post-test interview with examinees to help establish whether the target words were actually known or not, and we compared these results with the examinees' responses on the Levels Test.

One of the standard techniques in examining both tests and individual items is the comparison of responses to an individual item with scores on the overall test. This technique forms the basis of a number of procedures, including point-biserial coefficients and discrimination indices such as E_{1-3}. Where a number of items address the same underlying construct, then these procedures should work well because all the items are intended to relate to the construct in the same way. But vocabulary items are discrete items. Of course they may be added together to form an estimate of vocabulary size, but just because one target word is known does not necessarily mean that another will be. Even if most words in a frequency level are known, this does not guarantee that any particular word in that level will be; the fact that this is not the case is precisely why we test vocabulary in the first place. In addition, words are likely to have varying degrees of difficulty for learners from different first languages (Laufer, 1997), even if they are taken from the same frequency level. Thus, although it makes sense to select words in a level according to frequency criteria, it does not make sense to judge an individual item's validity according to how well other items (words) at the same level are answered. In short, item/global comparisons can be informative in drawing a test-writer's attention to potentially problematic items, but may not be the best means of establishing the soundness of those items.

Rasch analysis has become a standard way of analysing language

tests. However, one of the assumptions that must be met is that of 'local independence'. With the Levels Test, this raises the issue of whether the individual items within each cluster or 'testlet' are independent. Strictly speaking, the vocabulary items presented together in clusters cannot be considered independent. However, we feel that independence/nonindependence is unlikely to be a dichotomy, but is more likely to exist along a continuum. This is because independence stems not only from the item format itself, but also from the examinees' test-taking behaviour. Retrospective protocols (see Interview section below) indicate that when examinees know a target word, they usually answer the item directly without considering alternative options. On the other hand, if an examinee does not know the word or is uncertain, all of the options are generally reviewed. This means that if the words in a cluster are known, they are answered in an essentially independent manner. If one or more are not known, they will be dependent to varying degrees. For examinees with higher vocabulary sizes, the test overall will tend towards more independent behaviour.

This implies that examinees with lower vocabulary sizes will induce a tendency towards more dependence in the Levels Test. However, this may not always be the case. The rubric of the test discourages examinees from guessing blindly. As a result, in our study, we found that examinees generally left items blank if they did not know the answers. We found it very rare for examinees to choose the same distractor for all three answers in a cluster in order to improve the odds on answering items correctly. In cases where only 'known' words were attempted, the trend was towards relative independence, as above. In sum, although it is virtually impossible to determine precisely, we would cautiously suggest that there is at least a degree of independence within the clusters. Whether this is enough to meet the assumptions of Rasch analysis is debatable. An initial Rasch analysis of the tests' performance looking at the standardized residual correlations (using BIGSTEPS Table 10.6, Wright and Linacre, 1998) suggests that there is no evidence of dependence among the items in a cluster, but this needs to be investigated further.

Vocabulary knowledge is many-faceted (Richards, 1976; Schmitt, 1998; Nation, 2001), and no vocabulary test format currently available is able to tap into all forms of lexical knowledge. Likewise, it is difficult even to measure the degree of knowledge of single types of word knowledge confidently (e.g., meaning, collocation, appropriateness). Tests which attempt to do this (depth of knowledge tests) are currently being researched, but the Levels Test should be seen as a breadth of knowledge test (vocabulary size) with a much

more modest aim. (For more on tests of vocabulary size and vocabulary depth, see Read, 2000.) Because the words within a cluster have very different meanings, even a small amount of knowledge about a target word's meaning should enable a student to make a correct response. The Levels Test should, therefore, be seen as providing an indication of whether examinees have an initial knowledge of the most frequent meaning sense of each word in the test. (Many of the target words are polysemous, and this test does not address other meaning senses.) In addition, it is a receptive test and, as such, does not provide direct information about the ability to use the target words productively.

V The validation study

1 The preliminary stage

The first step was to investigate the item discrimination and test reliability indices of the existing Versions A–D when they were given to learners from a variety of backgrounds. Versions A and B were combined into a single test (Test E) and Versions C and D into Test F. Within Test E, Versions A and B were counterbalanced to create four variants, in order to control for any order effect. The same was done for Test F. Test E was given to a mixed group of 106 international students studying at a British university, with Test F being given to the same students one week later. The majority of students came from France (29), Germany (15), Spain (15), Malaysia (13), Japan (10) and China (7), with the rest coming from 13 other countries. The results were then analysed using ITEMAN (1989). In particular, we looked for clusters where items had distractors which attracted too many responses. Because we could not be sure that changing single items within a cluster would not affect the behaviour of the other items in the cluster, any poor item resulted in the complete cluster being discarded. The Cronbach alpha reliability figures (for dichotomously scored items) suggested that 10 clusters (30 items) per level would produce reliability figures above .90. (The original Levels Test and Versions A–D had only 6 clusters [18 items] per level). Because we discarded poor clusters and needed more items per version, we decided to combine the well-performing clusters and develop only two revised versions for final validation. The revised forms of Versions E and F were called Versions 1 and 2 respectively.

There were two main differences between these revised versions and Versions E and F. First, at the 2000 level, both new versions have 28 words from the first 1000 frequency level and 32 from the second 1000 level, so the proportion is closer to 1:1 than the 1:2 ratio

of the earlier versions. Second, since Versions A–D were written, an improved listing of academic vocabulary has been compiled from a new carefully-balanced academic corpus, the Academic Word List (AWL; Coxhead, 1998, 2000). The AWL has the advantage of giving better coverage of academic texts whilst listing fewer words than the University Word List (UWL; Xue and Nation, 1984). Rather than use the academic sections from the older versions, which were based on the outdated UWL, the lead author wrote new academic sections based on the AWL. Because the main study would be the first time these new sections could be analysed, a total of 24 clusters were written to allow for the discarding of poor clusters.

The facility values for the three items in each cluster were averaged to obtain a difficulty figure for each cluster. After the first trial the clusters were allotted to Versions 1 and 2 in a way which, we hoped, made the two versions equivalent in terms of difficulty. (Because changing any definition or target word in a cluster might have an effect on the behaviour of the other definitions or target words in that cluster, we worked once again with whole clusters rather than trying to move individual items between clusters.) At this point, we had two new versions of the test. There were 10 clusters in each section with the exception of the Academic section which had 12. For the final trial, the two versions were combined into a single instrument, which had two counterbalanced variants. It was then administered to the subjects. Once the results were analysed, we swapped one or two clusters per section between the two versions in order that the two versions should be of the same level of difficulty. In addition, four clusters were discarded from the Academic section, so that it now contained 10 clusters per version. Versions 1 and 2 of the Levels Test were then in their final form. (See Appendixes 1 and 2 for full renderings of Version 2; Version 1 is available in Schmitt, 2000.)

2 Subjects

Because the test population for the Levels Test is mainly envisaged as being learners of English for general or academic purposes, it was important to explore the test's behaviour with examinees from a variety of first languages and cultures. (A preliminary validation study by Beglar and Hunt (1999) included only Japanese subjects.) Overall, the researchers attempted to gather a sample population of learners of general and academic English that was as large and diverse as possible, even though this meant that the sample was not balanced. There was no attempt to select students according to their background knowledge, since Clapham (1996) has shown how difficult, if not impossible, it would be to do this. A total of 801 subjects were tested

in 13 groups at three sites in England, two in New Zealand, one in Slovakia, one in Brazil and two in Spain (see Table 1). Each subject was given all items from both Versions 1 and 2, with the exception of 56 subjects in Group 3, who took only the Academic sections. The subjects came from a variety of countries: Spain (322), Taiwan (142), Slovakia (103), Japan (68), Yugoslavia (40), countries in Southeast Asia (39), China (15), Brazil (12); the remaining 60 subjects came from 20 different countries. All subjects were learning English for general purposes or academic purposes, often with the goal of study at an English-medium university. Because the intended test population of the Levels Test can vary widely in proficiency, it was desirable to include a wide spectrum of proficiency levels in the sample population. Judging by the resulting vocabulary size scores, this goal was largely achieved.

It is possible that the Levels Test is also suitable for other populations than those explored here. A study is currently being planned to examine its use with young teenage learners of English as an additional language (EAL) in the British school system (Schmitt and Cameron, in preparation). As validation is an ongoing process, such

Table 1 Description of subject groups

Group	*n*	First language	Location	Learning context / purpose of English
1	192	mixed	England 1	General English summer school
2	22	mixed	England 2	MA–ELT course
3	64	mixed/ Japanese	England 2	Pre-sessional course preparing for entrance to English-medium university
4	18	mixed	England 3	MA-level language testing course
5	57	mixed	New Zealand 1	EOP / pre-sessional courses preparing for entrance to English-medium universities
6	29	mixed	New Zealand 2	EAP / pre-sessional courses preparing for entrance to English-medium universities
7	102	Slovak	Slovakia	Foreign language at secondary school
8	11	Portuguese	Brazil	Large private language school
9	98	Spanish	Spain 1	1st year university: general English
10	68	Spanish	Spain 2	1st year university: English translation
11	50	Spanish	Spain 2	2nd year university: English translation
12	56	Spanish	Spain 2	3rd year university: English translation
13	34	Spanish	Spain 2	4th year university: English translation

subsequent studies can further explore appropriate test populations and provide additional evidence about the test's characteristics.

VI Validity

Current views of test validity tend to conceive of it as a unitary notion encapsulating numerous aspects which contribute to acceptable test behaviour (Messick, 1989). However, in order to make this account as clear as possible, this validity section will present the various aspects of validity as separate entities.

1 Native speaker results

An initial requirement of most L2 tests is that they must be answerable by persons proficient in the language (see Davies *et al*., 1999). In this case we used native speakers to explore whether there were any indications that proficient English speakers would have problems with the Levels Test. Nation (1990) reports a native speaker subject achieving 100% on the original version of the Levels Test. For this study, nine native speakers (four British undergraduates and five postgraduates) took Versions 1 and 2. Their scores ranged from 307 to 312 with a mean of 309 (the maximum score was 312). Thus the Levels Test format appeared to pose no problems for these L1 speakers; all of them reached maximum or near-maximum scores.

2 Item analysis

The results from the 801 subjects were analysed using ITEMAN (1989), and each cluster's behaviour was explored. Although we suspect that the items in each cluster are not independent of each other, Rasch analysis suggests that the items do perform independently of one another, so we calculated the facility and discrimination indices (point-biserial) for each item. Table 2 gives the mean facility and discrimination indices for each of the levels. It can be seen that the mean facility value decreases as the levels contain words that are progressively less frequent. To give some sense of this, a typical cluster at the 2000 level included *lovely*, *slight* and *popular* with facility values of .84, .60 and .86, respectively; a typical 5000 level cluster included *mound* (.44), *eve* (.70) and *cavalry* (.63); while a typical 10 000 level cluster contained *keg* (.15), *alabaster* (.35) and *rasp* (.20). The one section that is not frequency-based, academic words, has a relatively high mean facility value, which would place it between the 2000 and 3000 levels in terms of difficulty. This raises the interesting question of where to locate it in regard to the other

Table 2 Facility values and discrimination indices (point biserial) for Versions 1 and 2

	Number of items	Item Facility		Discrimination Index	
		M	sd	M	sd
Version 1					
2000	30	.783	.097	.534	.115
3000	30	.663	.146	.664	.106
5000	30	.579	.146	.699	.074
10 000	30	.289	.176	.509	.233
Academic	30	.754	.094	.519	.087
Version 2					
2000	30	.783	.089	.541	.118
3000	30	.664	.170	.635	.117
5000	30	.579	.156	.695	.105
10 000	30	.290	.165	.546	.223
Academic	30	.756	.108	.519	.074

Note: Number of Students: 745

levels when administering the test. This will be taken up in the next section.

The mean discrimination indices vary from .509 to .669. For levels other than the 10 000 level, no individual item had a discrimination index of less than .30. At the 10 000 level, however, the discrimination indices for individual items fell below .30 in 13 out 60 cases (30 items per version times 2 versions). These 13 cases represented the most difficult items, with the facility values ranging from .03 to .16 (M = .10). We would argue that the discrimination indices for the Levels Test are acceptable, bearing in mind that vocabulary is learned as individual units and that it is quite usual for less able learners to know a certain number of relatively low-frequency words, while more able learners typically have some gaps in higher-frequency vocabulary.

The test rubric encouraged subjects not to guess blindly, 'If you have no idea about the meaning of a word, do not guess. If you think you might know the meaning, then you should try to find the answer.' From the data, it appears that examinees generally complied with this instruction and, particularly in the case of the less frequent words, did not answer an item if they did not know it. As a consequence, the proportion of examinees attracted by each individual distractor was not high. The items were surveyed and any item with a distractor attracting more than 10% of the examinees was flagged for closer inspection; this turned out to be only 18 of 300 items (6%). It seems that examinees tended to choose the correct answer, or leave the item blank. For example, the items in the cluster illustrated had facility

values of .52 (circular shape), .36 (top of a mountain) and .76 (a long period of time), while the proportion of examinees leaving them blank were .33, .34 and .15 respectively. In contrast, among the 15 distractors for these three items (five distractors per item), only two attracted responses above .04 (.10 and .14 for *top of a mountain*). The mean proportion of examinees selecting the wrong alternatives was only .12. If examinees guess blindly, with this item format the chances of guessing a wrong alternative should be far greater than guessing the correct answer. From these figures, however, we see that the distractors as a group are not chosen to any great degree. From this, it seems logical to infer that the vast majority of examinees were not inclined to guess blindly. This suggests that guessing is not a serious problem with this format and that correct answers do reflect some underlying knowledge of the target word. (However, see the Interview with Examinees section below for further discussion about guessing.) Item analysis, however, cannot address the issue of examinees not answering an item even though they have partial knowledge of a word. We consider this issue in the Interview with Examinees section below.

3 Profile of the sections

Research has shown that, in general, learners acquire more frequently used words before they acquire less frequently used ones (see, for example, Nation, 1990). Thus we can partially estimate the validity of the Levels Test by establishing whether learners do better on the higher frequency sections than on the lower frequency ones. We found that from a total of 30 possible, the mean for the four frequency levels were 25.29 (sd 5.80) for the 2000 level, 21.39 (7.17) for the 3000 level, 18.66 (7.79) for the 5000 level and 9.34 (7.01) for the 10 000 level, with analysis of variance plus Scheffé tests showing that the differences were all statistically significant ($p < .001$).

Read (1988) found similar differences between the frequency levels and went on to test for implicational scaling between them. The different sections were 'highly scalable', with the coefficients of scalability for his two administrations being .90 and .84. This means that criterion mastery of a lower frequency level implied mastery of all higher frequency levels. Like Read, we carried out a Guttman scalability analysis (Hatch and Lazaraton, 1991), using a criterion of mastery of 26 out the 30 possible per level. (This figure was chosen to be as close as possible to Read's criterion of 16 out of 18.) Details are given in Table 3.

Hatch and Lazaraton suggest figures of $> .90$ for the Coefficient

Table 3 Results of a Guttman scalability analysis

	Version 1	Version 2
C_{rep}	0.993	0.995
MM_{rep}	0.768	0.775
Scalability	0.971	0.978

Note: Order of Levels in both cases was 2000, 3000, 5000, 10 000

of Reproducibility (C_{rep}) and $>$ .60 for Scalability as minima for implicational scaling to be established. From this, it is clear that the four frequency sections have a very high degree of scalability. In most cases, therefore, if an examinee reaches the criterion at one level, the teacher or researcher can be reasonably confident that the higher-frequency levels are known at least to criterion mastery as well. For example, if criterion mastery of the 5000 level of Version 1 is reached, this indicates that the 3000 and 2000 levels are reached as well. (There will, of course, be exceptions to this. For example, some students in specialized subject areas may not as yet have acquired a wide basic vocabulary; see above under *Issues in Vocabulary Testing*.)

Our scalability analysis looked only at the frequency-based levels, while Read included the Academic section in his profile research and found that it did not fit. (The sampling method for the Academic section on the original Levels Test made it logical to place this section after the 5000 level, and for Read to analyse it in this manner.) We would argue that the words on the Academic section (examinees' mean score 22.65) are different in kind from the other levels and should not be included in the profile comparison. The academic words come from the AWL, which was compiled according to criteria of coverage and range across a variety of academic texts. Frequency was included as part of the criteria, but was not the dominant consideration. Thus, the words in the Academic section are not primarily frequency driven, as the other sections are. In fact, the facility values of individual items and Rasch item difficulty figures suggest that the words in the academic level fit in a broad range between the 2000 level and the 10 000 level. If the Academic section had to be fitted somewhere between the frequency levels on the basis of the results from this particular group of examinees, the above mean scores would best place it between the 2000 and 3000 levels. The main explanation for the relative ease of these words is that most are based on Latin or Greek roots and affixes, and so are relatively easy for examinees from a Romance language background (see the Equivalence section

below). Examinees from non-Romance backgrounds find them much harder.

However, other issues also affect the best placement of the Academic section. Most teachers would naturally wish to focus on more 'basic' 3000 level words before moving on to academic vocabulary, which would suggest placing it after the 3000 level. The Academic section could also be placed at the end of the test, as it is different in kind from the frequency-based levels. But this would entail having relatively easy words coming after a difficult section (10 000 level) in which many examinees may have given up. The solution is not to consider the Academic section as fixed in placement, but flexible according to the demands of each testing situation. Unfortunately, once the Levels Test has been set down in print, many teachers will use it only in that form. Because we have to present the sections in some sequence, we have balanced level difficulty and pedagogic issues and opted to place the Academic section between the 3000 and 5000 sections.

4 Factor analysis

The underlying construct that the Levels Test attempts to tap into is initial receptive knowledge of the given meaning sense of the target words. Since this is a relatively discrete construct, one would ideally expect only one factor to appear in a factor analysis of any particular section. For a factor analysis of all of the sections together, one might expect loading on one major factor of vocabulary knowledge and lesser ones for the different sections. This is exactly what happens. When the results of Versions 1 and 2 of each level were submitted to a Principal Components analysis, only one factor per section emerged, with each version loading at the same degree of magnitude (Table 4). This is not surprising as the test's format with its short definitions should require very little grammatical knowledge or reading ability; virtually the only apparent linguistic feature it could address is vocabulary.

Table 4 Principal component analysis of matching sections

Sections	Loading on factor	Percentage of variation explained
2000 1 and 2	.980	96.0
3000 1 and 2	.981	96.2
5000 1 and 2	.975	95.0
10 000 1 and 2	.979	95.9
Academic 1 and 2	.989	97.9

When all of the sections were analysed together, two factors emerged (Table 5), representing 78.4% and 10.4% of the variance respectively. What these factors are is open to many interpretations, and can only be resolved by further research. Here we simply propose two different hypotheses: either Factor 1 is a vocabulary knowledge factor and Factor 2 a difficulty factor which differentiates between the various levels, or Factor 1 represents knowledge of higher frequency vocabulary and Factor 2 knowledge of lower. With either interpretation, the results seem to support the assertion that the Levels Test is unidimensional, with the key measured trait being vocabulary knowledge.

Previous work on validating the Levels Test format (Beglar and Hunt, 1999) used factor analysis to evaluate individual test items rather than a complete level section of the test. We think this is not a particularly appropriate procedure for two reasons. The first concerns the item/global distinction. If we run a factor analysis on individual items from the same section and find one that does not load satisfactorily on the main factor, then this means that the subjects' responses to that item are not similar to their responses to the other items in the section. This essentially relates to the difficulty of the item. An unusual loading can be useful in highlighting poorly-written items but, assuming an item is sound, difficulty has no real bearing on whether a word belongs in a certain section or not; this decision is made on the basis of frequency alone. One would expect that some words would be either more or less well known than others in a section, and this can lead to atypical loadings, regardless of how sound their corresponding items are. In essence, just because an item has a low loading does not mean it is a bad item.

A related reason concerns the way in which the dominant factor at any level should be visualized. If we were looking at the 2000 level,

Table 5 Varimax rotation of the Levels Test sections (eigenvalue 1.0)

	Factor 1	Factor 2
2000 1	.926	.204
2000 2	.924	.217
3000 1	.853	.428
3000 2	.866	.385
5000 1	.758	.560
5000 2	.688	.639
10 000 1	.261	.929
10 000 2	.290	.918
Academic 1	.714	.553
Academic 2	.711	.563

then the factor might be hypothesized as 'knowledge of words at the 2000 frequency level'. But if we are looking at the individual words within that level, the only construct which makes any sense is 'knowledge of that particular word's properties'. Thus we believe that factor analysis of individual items confounds the separate constructs concerning section and item. For these two reasons, therefore, we feel that factor analysis of individual items is of limited value in this type of study.

5 Reliability

The reliability indices (Cronbach's alpha) for all of the Levels sections are high as illustrated by Table 6, and are in line with the .94 and .91 figures reported by Read (1988) for the original Levels Test. This shows that 30 items per level provides good reliability. In fact, if a reliability level of .90 is considered satisfactory, then the Spearman–Brown prophecy formula indicates that 24 items (8 clusters) would be sufficient in the case of the 10 000 section of Version 1, the section with the lowest reliability figure. However, the Levels Test attempts to estimate knowledge of large numbers of words. Even the section of the test with the smallest population of words, the Academic section, attempts to estimate how many of the 570 words in the AWL are known. It is therefore important to have as high a sampling rate as possible, which leads us to suggest using the complete 30 item sections as they are presented. This is especially true because the Levels Test is efficient in terms of time required (see the Practicality section below). If a shorter test is desirable, it may be reasonable to exclude certain sections (such as the 10 000 section for beginning learners, or the 2000 and 3000 sections for advanced learners) rather than shortening any particular section. However, it must be remembered that if the test is shortened, the reliability index is likely to become lower. Of course if there is any reason to measure certain levels with

Table 6 Reliability of the levels sections (Cronbach alpha)

Level	Number of items per version	Version 1	Version 2
2000	30	.920	.922
3000	30	.929	.927
5000	30	.927	.927
10 000	30	.915	.924
Academic	30	.958	.960

an even higher degree of confidence, then the sections from both Versions 1 and 2 can be combined to make a longer test.

6 Practicality

Results from the students in this validation study indicate that the Levels Test can be completed in a reasonable amount of time. Students who were timed in this study averaged 31 minutes (range 15–60) to finish a single version of the test. The test has other features which bolster its usability: it is quick and easy to score, can be readily photocopied, needs no special equipment and gives a more complete picture of a learner's vocabulary than most other tests. In addition, the test is easily computerized. There is already a computerized form of the original Versions A–D, and plans are being developed to put the new Versions 1 and 2 into an electronic format. Altogether, the Levels Test seems to rate highly in terms of practicality.

7 Interview with examinees

The above discussion begins to build an argument for validity, but much of the evidence is indirect. Schmitt (1999), however, argues for a more direct approach to construct validation by exploring how closely responses on vocabulary tests match examinees' knowledge of the lexical aspect being measured. In the case of the Levels Test, to truly determine whether the items are valid or not, we must determine whether the meaning of the target words on the test are known. One way of confirming knowledge is to conduct an in-depth interview with examinees about their knowledge of the target words (Schmitt, 1999).

a Procedure: Each complete version of the Levels Test contains 150 items (5 sections times 30 items). One third of the items from each section of Version 1 were selected (50 in total), ensuring that a range of item difficulties were sampled, and that the 3:2:1 (noun:verb:adjective) ratio was maintained. The two raters (the first two authors) then agreed upon the required elements of each word's definition. Piloting indicated that the two raters could achieve a high level of agreement, and that the 50 interview items could be completed in between 20 and 40 minutes. Subsequently over the course of a week 22 examinees of mixed proficiencies and nationalities from two English universities (labelled as England 1 and 2) were given Version 1 of the Levels Test. The sample comprised 8 Japanese, 5 Thai, 4 Chinese, 2 Venezuelan, 1 Botswanan, 1 Omani and 1 Korean students who were on general English summer school courses, on EAP pre-sessional courses or on mainstream university courses. The

vocabulary profiles from these students ranged from 77% (2000), 33% (3000), 17% (5000), 8% (Academic), 0% (10 000) for the weakest student to 100% (2000), 100% (3000), 100% (5000), 100% (Academic), 80% (10 000) for the strongest student, with other students being broadly distributed between these extremes. The written tests were not marked until after the interviews to avoid rater bias. The interviews began immediately after students had taken the written test. The students were first asked whether the Levels Test they had just taken was a reasonable test of vocabulary and whether they had had any problems with the cluster format. They were then shown two new clusters – one easy and one difficult – and were asked, retrospectively, to describe the process they had gone through as they answered the items in the written test. It was felt that giving new clusters would help the interviewees think not only retrospectively, but also introspectively as they worked their way through the new test items. Finally, they were interviewed on their knowledge of the 50 target words. The examinees were given a list of the target words so that they could see the words in their written forms. They were then asked to define each word. If they were able to supply an acceptable description, this was taken to mean the word was known. If the examinee was unable to supply a description, he or she was given a sheet of paper with the correct definition from Version 1 and four distractors taken from Version 2. For example, the options for *summit* were:

24 summit
a. top of a mountain
b. musical instrument
c. loan to buy a house
d. day or night before a holiday
e. soldiers who fight from horses

If the subject was able to select the correct definition and state that they were not guessing blindly, then he or she was also marked as knowing the word. This procedure provided *knows/doesn't know* ratings which could be compared with the *correct/incorrect* scoring of the Levels Test.

b Retrospective protocols: When asked if the Levels Test was a 'good' test of vocabulary, all subjects answered affirmatively. While the interview sample population is admittedly small, and learners sometimes say what interviewers want to hear, the interviews uncovered no evidence to indicate that examinees feel the test is unreasonable. This apparent examinee acceptance of the test parallels the previous findings of Beglar and Hunt (1999). Only three students reported any problems, and these were mainly to do with the desire for more context in the definitions. However, an examination of their

completed tests revealed that two subjects performed quite well (high percentage scores on each section except the 10 000 level), while the weaker subject still obtained 80% correct on the 2000 level. This suggests that their perceived problems reflect more of a preference for context-dependent items than any real problem in handling context-independent items, such as those appearing in the Levels Test. (For more on the role of context in vocabulary items, see Read, 2000.)

The retrospective protocols found that examinees tended to work down through the three stems in a serial manner. They usually reviewed all of the option words in the cluster when answering difficult items but, for easy items, they usually focused on the correct option without considering the others. This is similar to the test-taking behaviour reported in Read (1988).

An interesting question to come out of the research is the apparent mismatch between the results from the item analysis, which suggested that guessing was not at all widespread, and the results from the interviews in which nearly 23% (5 out of the 22) of the interviewees reported guessing when they did not know the words. The item analysis data confirms that distractors were chosen to a somewhat greater extent for items on lower frequency levels than for higher frequency levels. This would be congruent with a somewhat greater degree of guessing, but even low frequency items reflected low levels of distractor choice. Overall one would expect more guessing at lower frequency levels (examinees are most likely to know high frequency words and thus guess less at these levels). However, even with these low frequency items, the 'correct' answer was still by far the most frequent option chosen; and the distractors were chosen infrequently. This argues for low levels of guessing overall on the test.

The mismatch can be squared to some degree if we appreciate that the interviewees only reported guessing when they did not know the words. Even relatively weak students typically know many of the words in the higher frequency levels, so guessing would presumably not become pronounced until the lower frequency sections. Unfortunately, we did not ask the interviewees the *degree* of their guessing, only whether they had guessed at all during the test. Guessing is less serious for low proficiency examinees because they were generally unsuccessful with their guesses, but is an important issue for higher proficiency examinees who were more successful. Unsurprisingly, higher proficiency students were better able to eliminate the distractors and guess correctly, while weaker students were generally unable to do this to any great degree. However, in order to eliminate the distractors, examinees must know those words. This means that in order to improve their chances of guessing an unknown target word, they must know some of the other words at the same level

(distractors) which are not being directly tested. So to the extent that successful guessing is a result of distractor elimination, this guessing also indirectly indicates knowledge of the vocabulary at the targeted frequency level. In sum, our cautious initial interpretation is that (1) the effects of successful guessing behaviour on this test are no greater than might be expected for any receptive format, and (2) further research is needed to clarify this interpretation.

c Demonstration of knowledge of the target words' meanings: Using the interview procedure, the raters could be relatively confident of the examinees' knowledge of the targeted meaning sense of the test words. This is reflected in a high inter-rater reliability figure of .949 (Phi, $p < .001$) for the two subjects assessed by both raters in the main study and .970 for the three subjects rated in the interview piloting. The correlation between the written test (correct versus incorrect response) and the interview (demonstration versus nondemonstration of knowledge of the targeted meaning sense) for the 50 items was .749 (Phi, $p < .001$). This indicates that the Levels Test gives a reasonable estimate of learners' vocabulary knowledge, as it has been defined for the Levels Test.

The correlation was not perfect, however, and it was interesting to look more deeply into the mismatches between the written test and the interview results. They can be summarized in the contingency table, Table 7. From this table it becomes evident that in about 10% (b and c) of the 1098 total cases, the Levels result did not indicate the examinees' true lexical knowledge, as indicated by the interview. The mismatches came in two types. The first, knowing a word but not matching the correct option on the Levels Test (b in Table 7), does not seem to be too much of a problem with an occurrence rate of only about 4%. This addresses the case of examinees not attempting an item when they should, which was raised in the Item analysis section above. The second, not knowing a word but still matching the correct Levels option (c in Table 7), occurred slightly more often: at about 6%.

The second type is worth discussing because it touches on the

Table 7 Comparison of interview results with levels results

		LEVELS TEST		
		Correct	Incorrect	
INTERVIEW	Knew	a 731	b 47	778
	Did not know	c 65	d 255	320
		796	302	1098

element of guessing. Although results from the main study indicated that blind guessing was not a major problem (see the Item analysis section), five of the interviewed examinees said that they had guessed when they did not know the words. Those who said they occasionally guessed increased their score by from three to six words which they did not know, while one student who filled in every blank gained nine words. The nonguessing students generally had from one to three words correct on the sample from the Levels Test which they did not know. As these figures come from a 50-word sample, they need to be multiplied by three to indicate what might happen on the complete 150-item test.

The interviews also suggested that many of the mismatches were not the result of guessing, but of partial lexical knowledge. For example, one interviewee believed that *collapse* meant 'to break'. While this was not a full enough understanding to warrant a 'know' rating in the interview, it did allow him to choose the correct option, 'fall down suddenly', on the Levels Test. Once this (as yet unknown amount of) partial lexical knowledge is parcelled out of the mismatches, the remainder attributed to guessing seems less likely to pose a serious threat to the test's performance. While guessing has always been problematic with receptive test formats, the Levels Test generally seems to have good characteristics in this regard. However, the interview results suggest that test invigilators need to highlight the rubric against blind guessing before administering the test to examinees. Beyond this, further research needs to be carried out into how partial lexical knowledge affects examinees' responses as opposed to pure guessing.

VII Equivalence

Although we have presented some initial validity evidence for the Level Test, there is still the question of whether the two versions are equivalent. Henning (1987) states that this requires the equivalence of three features: means, variance and covariance. Table 8 illustrates the values for these three features for Versions 1 and 2 when the entire sample population is included (the 56 subjects who completed only the Academic section are included only in that section's analysis). From the table we can see that neither the means nor the variance is different statistically for any level. In addition, the covariance figures are satisfactory, being above .90. (The Pearson Product Moment correlation between the two complete tests is .95.)

Supplementary evidence for equivalence comes from the Principal Component analysis above (Table 5) where, for each level, both Version 1 and Version 2 load on the same factor and at the same load

Table 8 Equivalence of sections of Versions 1 and 2

Level	Mean		Variance		Covariance (1 and 2[c])
	1	2[a]	1	2[b]	
2000	25.293	25.296	33.211	34.071	.920
3000	21.369	21.411	53.247	49.624	.925
5000	18.659	18.666	61.273	60.096	.901
10 000	9.345	9.350	46.893	51.588	.918
Academic	22.622	22.674	71.473	71.790	.958

Notes: a All sets of means $p > .50$ (paired t-test); b All sets of variance $p > .50$, except 10 000 level $p > .05$ (Levene Statistic); c All covariances $p < .001$ (Pearson's)

weightings. Thus, according to Henning's criteria, the two versions are equivalent for the sample population. It would be useful to use Rasch analysis to investigate the equivalence more closely.

Developing different forms of a test that are truly equivalent has always been difficult. To confirm the above results, we analysed a number of subsets of the data to see if Versions 1 and 2 were equivalent for different language groups. The first language can make a significant difference in how difficult words are to learn. This is particularly true when it comes to cognates – words in two languages which come from the same parent word. Speakers of Romance languages have a distinct advantage in this regard, for many English words can be guessed according to their similarity to Romance words. As we have seen, this is particularly true for words in the Academic section. The target words in the Levels Test were chosen on a basis of frequency, and it is possible that either Version 1 or 2 could, simply by chance, contain more cognates than the other. (There are 472 words of Latin and 45 words of Greek origin in the 570-word Academic Word List (91%)). We had a large number of Spanish/Portuguese examinees whom we could isolate in order to give an indication of this, so the first subset we looked at consisted of Romance speakers. The results from 317 (respondents from Groups 8–13) of these subjects were re-analysed separately. Paired t-tests showed that for Romance speakers the results of the 2000, 3000 and AWL sections of Version 1 were not statistically different from those of Version 2 ($p > .10$), but that the 5000 level ($p < .001$) and the 10 000 level ($p < .05$) were different (Table 9). Thus, strictly speaking, the 5000 and 10 000 levels cannot be considered equivalent for the Romance speakers.

We also analysed the non-Romance population. We found that the means were not statistically different for the 2000, 3000 and 10 000 levels (paired t-test; $p > .15$), but were for the 5000 and Academic

Table 9 Mean correct scores for subject subgroups

Level	Romance Speakers		Non-Romance Speakers		Slovaks	
	Version 1	Version 2	Version 1	Version 2	Version 1	Version 2
2000	25.79	25.79	24.76	24.93	28.93	28.89
	(4.19)	(4.15)	(7.02)	(6.72)	(1.37)	(1.28)
3000	22.54	22.34	20.61	20.59	25.90	24.39*
	(4.89)	(4.96)	(8.76)	(8.07)	(3.34)	(3.78)
5000	20.36	21.49*	17.77	16.20*	23.26	21.32*
	(5.16)	(4.75)	(9.04)	(8.91)	(4.80)	(4.86)
10 000	12.41	12.85*	6.83	7.00	9.70	9.62
	(4.47)	(5.20)	(7.41)	(7.36)	(6.87)	(6.44)
Academic	26.08	26.09	23.36	24.30*	28.60	30.75*
	(3.89)	(3.61)	(11.32)	(11.91)	(6.92)	(7.08)

Note: * $p < .05$ (paired t-test)

levels ($p < .001$). In addition, we found that the 3000 level had unequal variance (Levene statistic; $p < .05$).

Finally we analysed the Slovakian subjects from Group 7, because they were a large group with a homogeneous first language and learning context, and because they were secondary school students, and were thus younger than most of the other subjects. The analysis showed that only the 2000 and 10 000 levels had both equivalent means ($p > .70$) and equivalent variance ($p > .50$).

This leaves us with a somewhat complicated set of results. Although the two versions appeared equivalent with the whole sample population, in fact only the 2000 level seemed to be equivalent in all of the analyses, and that is probably because most students did very well on it. Our interpretation is that Versions 1 and 2 cannot be considered truly equivalent, but that they produce very similar results, as can be seen in Table 9, where the significantly-different mean scores are relatively close in real terms. Given the relatively small scale of the differences, the two versions can probably be used in programmes as alternate forms, as long as no high-stakes comparisons are drawn from a comparison between the two, and as long as the potential differences in scores between the two versions are kept in mind. An example of this usage would be if a teacher wished to gain a general idea of whether his or her class had made vocabulary gains over a year's study. Another would be to use the two versions in alternate years as a vocabulary diagnostic test at the beginning of a course. If, however, there is a need to measure vocabulary size change longitudinally with maximum precision, then the same version should be used for both administrations. This is particularly true if individual, rather than group, scores are of interest. Although the two versions often

did not show a statistical difference in terms of group scores, individuals usually varied somewhat between the two versions at every level.

VIII Conclusion

Older versions of the Vocabulary Levels Test have been in use for quite some time without the benefit of a study into their behaviour. This article has described the construction of new Versions 1 and 2, and has provided initial evidence that they can provide valid results and produce similar, if not truly equivalent, scores. Native speakers do well on the test. The individual items appear to work well, with most examinees answering either correctly or not at all. The test supplies a profile of vocabulary frequency levels which are highly scalable. Factor analysis suggests that the test is essentially unidimensional. Personal interviews indicate that examinees accept the test and that answers on the test do reflect underlying lexical knowledge.

However encouraging this initial evidence is, the Levels Test needs to be explored further. Not only would it be interesting to investigate further the question of whether the items within a cluster are genuinely independent, but we should also investigate whether the test is suitable for use with other groups than those described in this study. It would also be useful to find out more about how guessing affects results, especially for more proficient students. Because the rubric encourages test-takers to try answering when they think they know a word, even if they are not 100% sure, there is a certain amount of ambiguity surrounding the boundary line at which individual examinees will determine that their level of knowledge about a word is sufficient to distinguish it from a guess. Different examinees are likely to set this boundary at different levels of certainty, with some being under-cautious and others over-cautious. Therefore, perhaps the most interesting research direction is the role of partial knowledge in vocabulary testing. Vocabulary learning is incremental, and most learners' knowledge of most target L2 words is likely to be incomplete (Schmitt, 2000). This means that discovering how to capture and interpret partial lexical knowledge in vocabulary assessment is an essential element in the development of the next generation of vocabulary tests.

Acknowledgements

Thanks to Averil Coxhead, Andrea Flavel, Sueli Fidalgo, Carmen Pérez Basanta, Ma Jesus Blasco, Maria Calzada-Pérez, Danica Gondov, Martha Jones and Tessa Moore for agreeing to administer the Levels Test to their students. Alex Gilmore provided the data entry,

and Mike Linacre provided useful input on the issue of item interdependence. We are grateful to Rita Green, Paul Nation and John Read for helpful comments on earlier drafts of this article. Special thanks to Paul and Nitha Nation for their hospitality during the three weeks when the original Versions A, B, C and D were written.

References

Bauer, L. and **Nation, I.S.P.** 1993: Word families. *International Journal of Lexicography* 6, 1–27.

Beglar, D. and **Hunt, A.** 1999: Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing* 16, 131–62.

Campion, M.E. and **Elley, W.B.** 1971: *An academic vocabulary list*. Wellington: New Zealand Council for Educational Research.

Clapham, C. 1996: *The development of IELTS: A study of the effect of background knowledge on reading comprehension.* Cambridge: Cambridge University Press.

Coady J. and **Huckin, T.** 1997: *Second language vocabulary acquisition: a rationale for pedagogy*. Cambridge: Cambridge University Press.

Cobb, T. 1997: Is there any measurable learning from hands-on concordancing? *System* 25, 301–15.

Coxhead, A. 1998: *An academic word list*. Occasional Publication No. 18, LALS, Victoria University of Wellington, New Zealand.

—— 2000: A new academic word list. *TESOL Quarterly* 34, 213–38.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. and **McNamara, T.** 1999: *Dictionary of language testing*. Cambridge: Cambridge University Press.

Hatch, E. and **Lazaraton, A.** 1991: *The research manual*. New York: Newbury House.

Hazenberg, S. and **Hulstijn, J.** 1996: Defining a minimal receptive second-language vocabulary for non-native university students: an empirical investigation. *Applied Linguistics* 17, 145–63.

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

ITEMAN 1989: St. Paul, MN: Assessment Systems Corporation.

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

Laufer, B. 1997: What's in a word that makes it hard or easy: some intralexical factors that affect the learning of words. In Schmitt, N. and McCarthy, M., editors, *Vocabulary: description, acquisition and pedagogy*: Cambridge: Cambridge University Press, 140–55.

Laufer, B. and **Nation, P.** 1999: A vocabulary-size test of controlled productive ability. *Language Testing* 16, 33–51.

Laufer, B. and **Paribakht, T.S.** 1998: The relationship between passive and active vocabularies: effects of language learning context. *Language Learning* 48, 365–91.

Meara, P.M. and **Jones, G.** 1990: *The Eurocentres Vocabulary Size Test 10K*. Zurich: Eurocentres.

Messick, S.A. 1989: Validity. In Linn, R.L., editor, *Educational measurement*. 3rd edn. New York: American Council on Education/Macmillan Publishing Company, 13–103.

Nagy, W.E., Herman, P.A. and **Anderson, R.C.** 1985: Learning words from context. *Reading Research Quarterly* 20, 223–53.

Nation, P. 1983: Testing and teaching vocabulary. *Guidelines* 5, 12–25.

—— 1990: *Teaching and learning vocabulary*. Boston, MA: Heinle and Heinle.

—— 2001: *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, P. and **Waring, R.** 1997: Vocabulary size, text coverage, and word lists. In Schmitt, N. and McCarthy, M., editors, *Vocabulary: description, acquisition, and pedagogy*. Cambridge: Cambridge University Press.

Read, J. 1988: Measuring the vocabulary knowledge of second language learners. *RELC Journal* 19, 12–25.

—— 2000: *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Richards, J.C. 1976: The role of vocabulary teaching. *TESOL Quarterly* 10, 77–89.

Schmitt, N. 1998: Tracking the incremental acquisition of second language vocabulary: a longitudinal study. *Language Learning* 48, 281–317.

—— 1999: The relationship between TOEFL vocabulary items and meaning, association, collocation, and word-class knowledge. *Language Testing* 16, 189–216.

—— 2000: *Vocabulary in Language Teaching*. Cambridge: Cambridge University Press.

Schmitt, N. and **Cameron, L.** in preparation: Vocabulary size and demands in English as additional language pupils at KS3/4.

Schmitt, N. and **McCarthy, M.**, editors, 1997: *Vocabulary: description, acquisition and pedagogy*. Cambridge: Cambridge University Press.

Schmitt, N. and **Meara, P.** 1997: Researching vocabulary through a word knowledge framework: word associations and verbal suffixes. *Studies in Second Language Acquisition* 19, 17–36.

Schonell, F. Meddleton, I., Shaw, B., Routh, M., Popham, D., Gill, G., Mackrell, G. and **Stephens, C.** 1956: *A Study of the oral vocabulary of adults*. Brisbane and London: University of Queensland Press / University of London Press.

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

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

Wright, B. and **Linacre, J.** 1998: BIGSTEPS: Rasch model computer program. Chicago, IL: Mesa Press.

Xue, G. and **Nation, I.S.P.** 1984: A university word list. *Language Learning and Communication* 3: 215–29.

Appendix 1 Student instruction sheet for the Levels Test

This is a vocabulary test. You must choose the right word to go with each meaning. Write the number of that word next to its meaning. Here is an example.

1	business	
2	clock	_____ part of a house
3	horse	_____ animal with four legs
4	pencil	_____ something used for writing
5	shoe	
6	wall	

You answer it in the following way.

1	business	
2	clock	6 part of a house
3	horse	3 animal with four legs
4	pencil	4 something used for writing
5	shoe	
6	wall	

Some words are in the test to make it more difficult. You do not have to find a meaning for these words. In the example above, these words are *business*, *clock* and *shoe*.

If you have no idea about the meaning of a word, do not guess. But if you think you might know the meaning, then you should try to find the answer.

Appendix 2 The Vocabulary Levels Test: Version 2 (© Norbert Schmitt)

The 2000 word level

1	copy	_____ end or highest point	1	accident	_____ loud deep sound
2	event		2	debt	
3	motor	_____ this moves a car	3	fortune	_____ something you must pay
4	pity		4	pride	
5	profit	_____ thing made to be like another	5	roar	_____ having a high opinion of yourself
6	tip		6	thread	

1	coffee	_____ money for work	1	arrange	_____ grow
2	disease		2	develop	_____ put in order
3	justice	_____ a piece of clothing	3	lean	_____ like more than something else
4	skirt		4	owe	
5	stage	_____ using the law in the right way	5	prefer	
6	wage		6	seize	

1	clerk	_____ a drink	1	blame	_____ make
2	frame	_____ office worker	2	elect	_____ choose by voting
3	noise	_____ unwanted sound	3	jump	
4	respect		4	threaten	_____ become like water
5	theater		5	melt	
6	wine		6	manufacture	

1	dozen	_____ chance	1	ancient	_____ not easy
2	empire	_____ twelve	2	curious	_____ very old
3	gift	_____ money paid to the government	3	difficult	_____ related to God
4	tax		4	entire	
5	relief		5	holy	
6	opportunity		5	social	

1	admire	_____ make wider or longer	1	slight	_____ beautiful
2	complain		2	bitter	_____ small
3	fix	_____ bring in for the first time	3	lovely	_____ liked by many people
4	hire		4	merry	
5	introduce	_____ have a high opinion of someone	5	popular	
6	stretch		6	independent	

The 3000 word level

1	bull	_____ formal and serious manner	1	muscle	_____ advice
2	champion		2	counsel	_____ a place covered with grass
3	dignity		3	factor	
4	hell	_____ winner of a sporting event	4	hen	
5	museum		5	lawn	_____ female chicken
6	solution	_____ building where valuable objects are shown	6	atmosphere	
			5	lawn	
			6	atmosphere	

1	blanket	_____	holiday
2	contest	_____	good quality
3	generation	_____	wool covering
4	merit		used on
5	plot		beds
6	vacation		

1	abandon	_____	live in a place
2	dwell	_____	follow in
3	oblige		order to catch
4	pursue	_____	leave
5	quote		something
6	resolve		permanently

1	comment	_____	long formal
2	gown		dress
3	import	_____	goods from a
4	nerve		foreign
5	pasture		country
6	tradition	_____	part of the body which carries feeling

1	assemble	_____	look closely
2	attach	_____	stop doing
3	peer		something
4	quit	_____	cry out loudly
5	scream		in fear
6	toss		

1	pond	_____	group of
2	angel		animals
3	frost	_____	spirit who
4	herd		serves God
5	fort	_____	managing
6	administration		business and affairs

1	drift	_____	suffer
2	endure		patiently
3	grasp	_____	join wool
4	knit		threads
5	register		together
6	tumble	_____	hold firmly with your hands

1	brilliant	_____	thin
2	distinct	_____	steady
3	magic	_____	without
4	naked		clothes
5	slender		
6	stable		

1	aware	_____	usual
2	blank	_____	best or most
3	desperate		important
4	normal	_____	knowing what
5	striking		is happening
6	supreme		

Academic Vocabulary

1	area	_____	written
2	contract		agreement
3	definition	_____	way of doing
4	evidence		something
5	method	_____	reason for
6	role		believing something is or is not true

1	adult	_____	end
2	vehicle	_____	machine used
3	exploitation		to move
4	infrastructure		people or
5	termination		goods
6	schedule	_____	list of things to do at certain times

1	debate	____	plan
2	exposure	____	choice
3	integration	____	joining
4	option		something
5	scheme		into a
6	stability		whole

1	alter	____	change
2	coincide	____	say something
3	deny		is not true
4	devote	____	describe
5	release		clearly and
6	specify		exactly

1	access	____	male or
2	gender		female
3	psychology	____	study of the
4	license		mind
5	orientation	____	entrance or
6	implementation		way in

1	correspond	____	keep
2	diminish	____	match or be
3	emerge		in agreement
4	highlight		with
5	invoke	____	give special
6	retain		attention
			to something

1	edition	____	collecting
2	accumulation		things over
3	guarantee		time
4	media	____	promise to
5	motivation		repair a
6	phenomenon		broken
			product
		____	feeling a
			strong reason
			or
			need to do
			something

1	bond	____	make smaller
2	channel	____	guess the
3	estimate		number or
4	identify		size of
5	mediate		something
6	minimize	____	recognizing
			and naming
			a person or
			thing

1	explicit	____	last
2	final	____	stiff
3	negative	____	meaning
4	professional		‘no’ or ‘not’
5	rigid		
6	sole		

1	abstract	____	next to
2	adjacent	____	added to
3	neutral	____	concerning
4	global		the whole
5	controversial		world
6	supplementary		

The 5000 word level

1	analysis	____	eagerness
2	curb	____	loan to buy a
3	gravel		house
4	mortgage	____	small
5	scar		stones
6	zeal		mixed with
			sand

1	artillery	____	a kind of tree
2	creed		
3	hydrogen	____	system of
4	maple		belief
5	pork	____	large gun
6	streak		on wheels

1	cavalry	_____	small hill
2	eve	_____	day or night before a holiday
3	ham		
4	mound		
5	steak	_____	soldiers who fight from horses
6	switch		

1	chart	_____	map
2	forge	_____	large beautiful house
3	mansion		
4	outfit	_____	place where metals are made and shaped
5	sample		
6	volunteer		

1	circus	_____	musical instrument
2	jungle		
3	trumpet	_____	seat without a back or arms
4	sermon		
5	stool		
6	nomination	_____	speech given by a priest in a church

1	revive	_____	think about deeply
2	extract		
3	gamble	_____	bring back to health
4	launch		
5	provoke	_____	make someone angry
6	contemplate		

1	shatter	_____	have a rest
2	embarrass	_____	break suddenly into small pieces
3	heave		
4	obscure		
5	demonstrate		
6	relax	_____	make someone feel shy or nervous

1	decent	_____	weak
2	frail	_____	concerning a city
3	harsh		
4	incredible	_____	difficult to believe
5	municipal		
6	specific		

1	correspond	_____	exchange letters
2	embroider		
3	lurk	_____	hide and wait for someone
4	penetrate		
5	prescribe	_____	feel angry about something
6	resent		

1	adequate	_____	enough
2	internal	_____	fully grown
3	mature	_____	alone away from other things
4	profound		
5	solitary		
6	tragic		

The 10 000 word level

1	alabaster	_____	small barrel
2	tentacle	_____	soft white stone
3	dogma	_____	tool for shaping wood
4	keg		
5	rasp		
6	chandelier		

1	throttle	_____	kindness
2	convoy	_____	set of musical notes
3	lien	_____	speed control for an engine
4	octave		
5	stint		
6	benevolence		

1	bourgeois	_____	middle class people
2	brocade	_____	row or level of something
3	consonant	_____	cloth with a pattern or gold or silver threads
4	prelude		
5	stupor		
6	tier		

1	scrawl	_____	write carelessly
2	cringe	_____	move back because of fear
3	immerse	_____	put something under water
4	peek		
5	contaminate		
6	relay		

1	alcove	_____	priest
2	impetus	_____	release from prison early
3	maggot	_____	medicine to put on wounds
4	parole		
5	salve		
6	vicar		

1	blurt	_____	walk in a proud way
2	dabble	_____	kill by squeezing someone's throat
3	dent	_____	say suddenly without thinking
4	pacify		
5	strangle		
6	swagger		

1	alkali	_____	light joking talk
2	banter	_____	a rank of British nobility
3	coop	_____	picture made of small pieces of glass or stone
4	mosaic		
5	stealth		
6	viscount		

1	illicit	_____	immense
2	lewd	_____	against the law
3	mammoth	_____	wanting revenge
4	slick		
5	temporal		
6	vindictive		

1	dissipate	_____	steal
2	flaunt	_____	scatter or
3	impede		vanish
4	loot	_____	twist the
5	squirm		body about
6	vie		uncomfortably

1	indolent	_____	lazy
2	nocturnal	_____	no longer
3	obsolete		used
4	torrid	_____	clever and
5	translucent		tricky
6	wily		