Testing L2 Vocabulary: Current Test Formats in English as a L2 Used at Swedish Universities

HENRIK GYLLSTAD

Abstract
Recent literature on L2 vocabulary testing points to a move away from discrete vocabulary testing towards more embedded and integrative approaches. This paper investigates if the way vocabulary is currently tested at English departments at Swedish universities follows this trend. The survey shows that a great majority of universities use discrete and selective vocabulary tests and that the multiple-choice format is the most widely used format. The paper takes a closer look at this format, in particular a 120-item version used by a great number of departments. A small-scale study of the results on this test, by university learners of English in their first, second and third term of study, is presented. The study shows that learners on the most advanced level of study scored significantly better than the two less advanced levels, but that no significant difference could be observed between scores of first term and second term full-time learners. It also shows that learners' scores are highly scalable, lending validation support to the underlying assumption of the test design.

1 Introduction

Recent authoritative literature on second language (L2) vocabulary acquisition and testing (Read 2000), traces a shift in Western education from discrete-item vocabulary tests to more comprehensive and embedded tests, where vocabulary is measured as part of overall language proficiency. This change can be seen as a tangible effect of the advent in the late 20th century of a more communicative approach to language learning that superseded the more structural approaches prevailing before (cf. Schmitt 2000). At the same time, since communicative approaches are predominantly oral, and oral language is more restricted in range than written language, it might be the case that less emphasis is put on vocabulary on the whole compared to structural approaches[1]. This paper investigates if a move away from discrete-item vocabulary testing is visible in reality in the case of English as a L2 on university level in Sweden today. The overall question it sets out to answer is if English vocabulary is tested as an integral part of language proficiency or if it is tested as a discrete construct of its own.

The outline of the paper is as follows; in section 2, two influential distinctions widely used in L2 vocabulary testing are discussed. In section 3, the survey of what English vocabulary tests are presently used at Swedish universities is

[1] I would like to thank Paul Meara (p.c.) for pointing this out.

presented. Section 4 provides a closer look at the most commonly used vocabulary test format according to the survey: a 120-item multiple-choice test. Finally, section 5 reports a small-scale study of how learners from three different learner levels performed on this test, finally, section 6 sums up the findings of the paper.

2 Two influential distinctions in L2 vocabulary testing

Certain widespread assumptions seem to govern the field of L2 vocabulary testing. Before we look into how English vocabulary is tested today at Swedish universities, it is relevant to address some of the most commonly used assumptions, since they will be part of our subsequent analysis and discussion of test methods and designs. Consequently, this section serves to highlight two apparent dichotomies which recur in vocabulary testing research. These are, on the one hand, the distinction between vocabulary breadth and depth, and on the other, receptive and productive knowledge. I will below discuss these two. Essentially, the two distinctions have to do with different constructs of word knowledge. I here follow Chapelle's definition of the term construct to mean "a meaningful interpretation of observed [language] behaviour" (1998:33).

2.1 Breadth and depth

The terminology used in the literature may be somewhat confusing. In the literature, 'breadth' is used interchangeably with 'size', and 'depth' is sometimes substituted for 'organisation' or 'quality'. In this paper, I will use 'size' and 'depth' since I think that these two terms better reflect the aspects in question. Size, then, will be used to designate how many words a learner knows, whereas depth will be used to designate different aspects of how well a learner knows a word.

The size of a person's vocabulary, firstly, is a construct that has received a lot of attention by researchers. Quite a few studies have been conducted with the aim of trying to estimate the size of a learner's vocabulary. Basically, there are two conventionalized ways of going about this. One way is to take a sample from a dictionary and the other is to use a sample from a frequency list based on a corpus. The dictionary-based technique implies that a representative sample of words (every *n*-th word) is taken from the dictionary and that the native speaker is tested on those words (see Nation 1993). The rationale behind this is that the score on the test may be generalized to the total number of words in the dictionary[2]. For example, if the sample consisted of one in every 10 words in the

[2] Employing a so-called spaced sampling for test purposes may lead to a sampling problem. If, for example, the first word on every fifth page is used, then due to the fact that high-frequency words have more entries per word and more spacious entries in the dictionary, the result will be that more high-frequency words will end up in the sample than there should be.

sample, then the test-taker's scores on the test would be multiplied by 10 to get the overall vocabulary size. Examples of this approach can be found in Goulden et al. (1990) and D'Anna et al. (1991), who focused on native speakers. The technique used for the compilation of a frequency list is intrinsically based on some sort of corpus. The corpus may either be a general corpus or a specialized one. An example of a frequency list based on a specialised corpus is *The Academic Word List* (Coxhead 1998, 2000), and examples of well-known and commonly used frequency lists based on more general corpora are *The Teacher's Word Book*[3] (Thorndike and Lorge 1944), *The General Service List*[4] (West 1953) and a list based on the Brown corpus, provided by Francis and Kučera (1982). Normally, the words of frequency lists are arranged in different bands: the 1,000 most frequent words, the second thousand most frequent words, etc, and tests based on these types of bands are designed on the same assumption as the dictionary-based ones: if a test taker knows a proportion of the sample items from a particular band, then we can generally assume that she will know a similar proportion of all the words in that band.

Secondly, as opposed to vocabulary size, which gives a rather superficial indication of word knowledge, the concept of vocabulary depth refers to various aspects of how well a word is known. Anderson and Freebody (1981: 92) described it in relation to what would be understood by an ordinary [native speaker] adult under normal circumstances. A person, they claim, can be seen as having a sufficiently deep understanding of a word if it conveys to her all distinctions that are available to the ordinary adult speaker. Compared to the supply of studies on vocabulary size, the concept of vocabulary depth is sparsely explored. The concept of depth is closely linked to the question of what it means to know a word. A number of researchers have over time tried to define, more or less exhaustively, what knowing a word entails. Most of these attempts result in listing numerous criteria (e.g. Cronbach 1942; Richards 1976 and Nation 1990). These can all be seen as more or less complementary to each other.

Paul Nation lists different aspects of vocabulary knowledge for testing in a table in his recent monograph (2001: 347). Nation's table is reproduced below as Table 1. Depth of word knowledge can be seen to involve, to varying extent, aspects from column 2, such as concept and referents, associations, grammatical functions, collocations and constraints on use. In comparison, vocabulary size tests generally tap learners for knowledge of form aspects (column 1) and form and meaning (column 2). Worth noting is the systematic distinction between receptive and productive knowledge. This distinction will be briefly discussed in the following subsection (2.2). Normally, tests of depth of word knowledge incorporate relatively few items since investigating depth is a complex venture.

[3] Contains about 13,000 word families based on an 18,000,000 million word written corpus.
[4] Contains 2000 headwords based on a 5,000,000 word written corpus.

More time is generally spent on each test item and consequently fewer items can be tested.

Form	spoken	R	Can the learner recognize the spoken form of the word?
		P	Can the learner pronounce the word correctly?
	written	R	Can the learner recognize the written form of the word?
		P	Can the learner spell and write the word?
	word parts	R	Can the learner recognize known parts in the word?
		P	Can the learner produce appropriate inflected and derived forms of the word?
Meaning	form and meaning	R	Can the learner recall the appropriate meaning for this word?
		P	Can the learner produce the appropriate word form to express this meaning?
	concept and referents	R	Can the learner understand a range of uses of the word and its central concepts?
		P	Can the learner use the word to refer to a range of items?
	associations	R	Can the learner produce common associations for this word?
		P	Can the learner recall this word when presented with related ideas?
Use	grammatical functions	R	Can the learner recognize correct uses of the word in context?
		P	Can the learner use this word in the correct grammatical patterns?
	collocations	R	Can the learner recognize appropriate collocations?
		P	Can the learner produce the word with appropriate collocations?
	constraints on use (register, frequency…)	R	Can the learner tell if the word is common, formal, infrequent, etc.?
		P	Can the learner use the word at appropriate times?

Note: In column 3, R = receptive knowledge, P = productive knowledge

Table 1. Aspects of vocabulary knowledge for testing, from Nation (2001:347).

Finally, as a matter of interest, Paul Meara (p.c.) advocates an alternative approach to vocabulary acquisition and the lexicon. Rather than seeing the concept of depth, or as he prefers to call it, organization, as something relating to individual words, Meara proposes a view in which organisation applies to the whole lexicon. Thus, in his view, a learner is not seen to have depth of word knowledge of certain words, but can rather be shown to have a vocabulary that is more structured, like a network with a high degree of connectivity between words in the lexicon. This view is based on experiments involving association tasks where L1 speakers were found to have more connections between words in the lexicon than did L2 speakers (see Meara 1996 and Meara and Wolter, in Press).

2.2 Receptive and productive knowledge

As is apparent in Table 1 above, it is customary for researchers to make use of a distinction between receptive and productive knowledge of vocabulary items.

References to this distinction are traced back to the middle of the 19th century (Waring 1999). In relation to vocabulary, Nation (2001: 24-25) defines receptive use as involving "perceiving the form of a word while listening or reading and retrieving its meaning", whereas productive use "involves wanting to express a meaning through speaking or writing and producing the appropriate spoken or written word form".

It is widely agreed that a language user, in general, can recognize and understand more words than she can use when speaking or writing. That a learner should use a word in production, and not be able to recognize or understand it receptively, I think, goes against common, and linguistic, sense. There has to be an initial exposure to a word involving listening or reading that precedes the first productive instance of it. However, it is of course conceivable that a learner's first receptive encounter with the word merely involves recognition of the form, spoken or written, and that any subsequent attempts to use it may be infelicitous due to lack of understanding of the proper meaning of the word. Conversely, a learner may use a word frequently when speaking to connote a specific concept, but could in theory fail to recognize the conventionalized orthographic representation denoting the concept. This might be more common in cases where the learner's L1 is very different from the L2, i.e. belonging to a different language family with few cognate words and different orthography and phonology. Take, for example, English words like *subpoena, mortgage,* and *wreath*. A person who has only heard these words may deny their existence as words of English in a recognition test like a checklist or yes/no test because of the discrepancy between pronunciation and spelling (see e.g. Meara 1996 for a brief account of this test format).

Research carried out on size differences between receptive and productive vocabulary of L2 learners (Waring 1997) has shown that learners scored better receptively than productively on a passive definition-matching test and a controlled active test. Also, the learners' receptive vocabulary became progressively larger than their productive vocabulary as their overall vocabulary size grew.

In terms of test design, it is of course crucial to decide what underlying construct is to be tested, and taking the distinction of receptive versus productive into account seems unavoidable. However, as with most other dichotomy-like phenomena, when put under the magnifying glass, it tends to lose its clear-cut nature. The distinction between receptive and productive vocabulary is no different, and researchers suggest different analyses.

A popular metaphor to use in these contexts is the continuum, allowing for gradual differences. Melka (1997) discusses degrees of familiarity a learner might have with a word, stating that phonological, morphological, syntactical and lexical information about an item constitutes a very high degree of familiarity, whereas merely having visual recognition ability suggests a low degree of familiarity. On the whole, Melka admits to the existence of empirical

evidence for a difference between receptive and productive vocabulary, but dismisses a proper dichotomy (ibid.:101), and suggests the use of a continuum with degrees of familiarity. Meara (1990) proposes a diverging view from that of Melka. Meara argues that active vocabulary may be seen as existing on a continuum, but that passive may not. The reason for this is that passive vocabulary may only be accessed by means of appropriate external stimulation. He claims that there are no internal links available between the 'passive' word and other words in the lexicon network. Furthermore, Read (2000: 154-157) calls for more narrow definitions of the terms production and reception in relation to testing purposes, introducing *recognition* and *recall*, and *comprehension* and *use*. Recognition is taken to involve tasks where a learner is supposed to show that she has understood the meaning of a target word presented to her. Recall involves the presentation of some sort of stimulus, based on which the learner is expected to recall the target word from memory. Comprehension and use are seen to involve more context-dependent and comprehensive measures. Comprehension involves a task where the learner must show whether she understands a word given in a context, whereas use is involved when the learner is asked to produce one or several words, for example in oral retellings, translations and picture description tasks.

Irrespective of Melka's and Read's elaborations, and despite Meara's proposal for a different analysis, the two-fold distinction seems to be a widely used, die-hard notion which to a great extent affects the thinking of test designers and L2 vocabulary researchers alike.

3 Vocabulary testing at English departments in Swedish universities

Taking into account the development described in the introduction, it is interesting to see how well this is reflected in the higher education system in Sweden when it comes to the subject of English. According to information provided by the National Agency of Higher Education, there are 39 higher education institutions in Sweden. Out of these 39 institutions, 25 offer English as a subject of study. As a first step, I contacted these 25 different departments and asked if and how they tested English vocabulary in their respective syllabi.

The information I obtained is listed in Table 2 below (universities are in alphabetical order). Some of the departments supplied sample copies of the tests they were using at the time of the investigation, which allowed for closer scrutiny of the designs, but in most cases, the information provided in the table was based on self-report from the various departments. As is quite clear, already from a quick glance at the table, a majority of the departments use a discrete vocabulary test. The term discrete vocabulary test refers to a test in which vocabulary is measured as an independent construct.

University (College)	Department/School	Use of discrete test		Format of discrete test/tests
		YES	NO	
Blekinge I.T.	Dept of Humanities	X		Multiple choice[5]
Dalarna U.C.	School of Arts and Edu.	X		Match L2 words w. L2 syn. + gap-filling choice fr fin. list)
Gothenburg U.	Dept of English	X		Multiple choice[5] + CMCT[6]
Karlstad U.	Dept of Culture and Comm.	X		Multiple choice[5] + CMCT[6]
Kristianstad U.C	Dept of Hum & Soc. Sciences	X		Multiple choice[5]+ L2>L1 transl.
Linköping U.	Dept of Language & Culture	X		Multiple choice[5]
Luleå U.	Dept of Comm. & Languages		X	
Lund U.	Dept of English	X		Multiple choice[5]
Mid Sweden U.C.	Dept of Humanities		X	
Stockholm S.E.	Dept of mod. Lang. and Hum.		X	
Stockholm U.	Dept of English	X		Multiple-choice + gap-filling w L1 cue
Södertörns U.C.	Dept of Lang. & Cult. Studies	X		Multiple choice
Umeå U.	Dept of English	X		Multiple-choice
U.C. Borås	School of Edu. and Behav. Sc.	X		Multiple choice[5] + CMCT[6]
U.C. Gävle	Dept. of Hum. and Social Sc.		X	
U.C.Halmstad	Dept of Humanities	X		L2>L1 transl. (w. in cont.) + Prod. L2 def. or syn. of L2 ws.
U.C.Jönköping	School of Edu. and Comm.	X		Multiple choice[5] + CMCT[6]
U.C Kalmar	Dept of Arts	X		Multiple-choice (L2 monolingual def.-matching)
U.C Malmö	Int. Migration and Ethnic Rel.		X	
U.C Mälardalen	Dept of Humanities		X	
U.C Skövde	Dept of Languages	X		Multiple choice[5] + CMCT[6]
U.C Trollhättan/ Uddevalla	Dept of Soc. & Behav. Studies	X		Multiple-choice
Uppsala U.	Dept of English	X		Multiple-choice (60 L2 target items)
Växjö U.	Dept of Humanities	X		L2>L1 transl. (word in context)
Örebro U.	Dept of Humanities	X		Multiple-choice + L2>L1 transl. (words in context)
N = 25	N = 25	19	6	

Table 2. Tests at Swedish universities, university colleges etc.

As can also be seen in the table, multiple-choice seems to be the predominating L2 vocabulary test format used in courses of English at Swedish universities today. Out of the total 25 departments consulted, 19 stated that they did use a specific and discrete vocabulary test of some kind. Furthermore, out of these 19 departments using a discrete vocabulary test, as many as 16 claimed that they used a multiple-choice format. This is an interesting finding since it largely goes against the credo of the communicative approach to language teaching which is said to have been the dominating one of the Western world during the last fifteen to twenty years (see Read 1997: 303, 2000: 3-5; Schmitt

[5] 120-item test described in section 4 of this paper

[6] Contextual Multiple Choice Test which to some extent taps aspects of vocabulary knowledge, e.g. collocations and idioms, but focuses on what is generally considered to be more grammatical structures (e.g. prepositions, verb paradigms, word order)

2000: 14). One explanation, perhaps, to the yet present and extensive use of the format is that a more structural approach to English language teaching and testing may still prevail, as it were, at a majority of the departments in Sweden. Objective testing, that is, testing where correct answers are clearly specified and markers are not required to make any judgements, has for a long time been part of the so-called *discrete-point approach* to language testing, which was in vogue in the latter half of the 20th century (see Read 2000: 77). In this approach, assessment was focused on individual structural elements of a language and learners' knowledge of these. Alternatively, the test situation reflected in Table 2 may harbour an adoption of a more communicative and holistic approach to teaching, but at least when it comes to vocabulary testing, more conservative methods may, in a manner of speaking, have survived[7].

A conclusion to be drawn from the survey is that 19 out of 25 departments find it worthwhile and meaningful to test vocabulary knowledge as an independent construct, as opposed to, or in combination with more comprehensive measures. It is important to note that a majority of the departments that said they were not using any discrete vocabulary tests emphasized that they preferred to assess vocabulary in a more holistic and embedded way, through, for example, essay writing and oral presentations.

Out of the 16 departments who said that they use a multiple-choice test format, 10 reported that they use a particular 120-item multiple-choice test originally designed by researchers at Gothenburg University in the late 1960s. Because of the wide popularity of this test, it would be worthwhile to take a closer look at it, and discuss the format and design in greater detail.

4 The Multiple-choice Format and The 120-item test of English

4.1 The multiple-choice format

The multiple-choice format is one of the most widely used formats in vocabulary testing (Read 2000: 77). It is not difficult to see why. It is generally considered to be easy to mark. On the minus side, however, can be said that it is quite difficult and time-consuming to construct. The format consists of a number of test items in which the test-taker is required to choose the correct option, called the *key*, from several alternatives provided. There are numerous sub-formats employed. A typical example of a multiple-choice test can be seen in the example below (Read 1997: 305)

[7] See Mobärg (1997) for a discussion on differences between structural and lexical approaches to L2 teaching and testing.

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

The item above is of a sentence completion type, and the test-taker is supposed to choose from the list of four alternatives and insert the one considered to be correct. Quite often, answer sheets are designed for computer scoring and statistical calculations, which lighten the burden on teachers and course administrators even more. Unless the targeted test items are given with a lot of context, the format allows for a high sample ratio. The alternatives, arranged in a random order, are made up by the key and a number of *distractors*. In some cases, richly contextual material provides the stimulus to the item, but the item can also appear in an isolated fashion. Since the format requires the test-taker to recognize a proper item among several possible ones, the test taps knowledge on the receptive side of a word knowledge continuum such as the one described in section 2.2 (suggested by Melka 1997).

Analyses of scores derived from multiple-choice tests normally give a clear picture of the reliability as well as difficulty of the test. Furthermore, distractors can be analysed with regard to how well they function. It has been shown that it is possible to regulate the level of difficulty by varying the closeness in meaning between the distractors and the key. The type of distractor found to be the most difficult is the 'false' synonym, that is, a synonym with a similar meaning to that of the key, but one which does not fit the context (Nation 2001: 349-350).

The guessing factor is a commonly addressed problem, especially if the test items include few alternatives. Also, the role of the distractors has been debated. As pointed out by Davies et al. (1999: 125), a test may assess test-takers' ability to reject obviously incorrect distractors rather than their actual knowledge of the target item.Despite the widespread popularity of the format in general, Read (2000:78) notes the paucity of ongoing research on multiple-choice tests applied to second language learning.

4.2 The 120-item multiple-choice test of English

In the previous section, we saw that a particular type of multiple-choice test of English as a second language is used extensively at Swedish universities today. For this reason, this test merits closer investigation with regard to origin, design, underlying assumptions and rationale.

Following the terminology of Read (2000: 9), the test in question is a discrete, selective and context-independent test. Discrete means that it measures vocabulary knowledge as an independent construct, selective means that it is a measure which focuses on specific vocabulary items, and context-independent

means that it is a measure in which the test-taker can produce the expected response without referring to any context. The test is essentially a measure of vocabulary size, and its principles were developed by Allvar Ellegård in the 1960s (see Ellegård 1960). Furthermore, Zettersten (1979) reports of experiments in which the test format was used to render cross-national comparisons of first-year university students' scores as a measure of vocabulary proficiency.

The test consists of 120 English words. Each word is given together with 5 Swedish words, out of which 1 is a translation equivalent, i.e. the key, and 4 are distractors. Each item is thus presented in the following way (English glossing within single quotation marks under each Swedish alternative and the asterisk indicating the key are my additions):

	A	B	C	D	E
ASSIST	förbättra	hjälpa*	hävda	motstå	praktisera
	'improve'	'help'	'assert'	'resist'	'practice'

The target item on the left is supposed to be matched with one of the five Swedish alternatives, A-E, to the right. Only one of them can be chosen as the correct answer. Originally, the test was not a multiple-choice design. Instead, the test-taker was asked to produce a synonym, a translation equivalent or a definition. However, from the late 1960s and onwards, the present design has been used (Mobärg 1997: 214).

The 120 test items are selected on the basis of frequency. See subsection 2.1 for a brief description of the techniques behind the compilation of frequency lists and frequency bands. The test consists of 6 parts of 20 items each, where each 20-item part corresponds to a particular frequency band. The underlying assumption is that the more frequent a word is in a language, the more probable it is that a learner knows it. For each part of the test, there is consequently an expected increase in difficulty. This assumption will be further investigated in section 5 below, where the scalability of the test will be evaluated through the analysis of test score data from 90 students from three learner levels. The word list employed in the creation of the test is essentially that of Thorndike and Lorge (1944), which gives frequency information largely based on literary texts, such as English classics, textbooks and popular magazines. However, based on that, Thorén (1967) subsequently compiled an adapted list tailored for different school stages of the Swedish education system, and it is the latter publication from which test items are drawn. As has been reported by Mobärg (1997), the vocabulary tested in the Gothenburg test was drawn from the approximately 10,000 most common words of English. However, the 700 or so most common words are not included in the test, and close English-Swedish cognates are also excluded.

The relation between the six parts, frequency bands and word list notation is presented in Table 3 below (adapted from Mobärg 1997: 214). The notation in the third column requires a brief explanation. The author of the wordlist aimed at defining what words should be learnt at the various stages in the education system. The notation stands for different classes or stages. Thus, in classes 7-9, corresponding to secondary school, where learners are approximately 13-15 years old, about 2000 words are earmarked for learning.

Test part	Number of words in frequency band	Notation in Thorén (1967)
1	2000	7, 8, 9
2	800	G pa and G1
3	800	G2
4	800	G3
5	1700	G+
6	2800	Gx

Table 3. Relation between the test parts of the 120-item multiple-choice test, number of words in each sampled frequency band and sampled word list.

Before this stage, Thorén provides a compilation of about 700 words for classes 4-6. The designations 'G pa and G1', 'G2' and 'G3' stand for the three stages of upper secondary school (*gymnasium*, in Swedish), where learners normally are between 16 and 18 years old. 'G+' stands for extra words for advanced students and 'Gx' for words to be acquired at Colleges of Education (Thorén 1967, English introduction). The increase in number of words in the frequency bands tested in parts 5 and 6 is simply due to the fact that low frequency bands contain more words.

A hypothetical estimation of vocabulary size can be made on the basis of a learner's score on the test. If a learner scores 15 out of 20 in part 1 of the test, the sample of which is taken from a band of about 2000 words, then that learner can be expected to know 75 per cent of the target words in that band, i.e. about 1500 words. Similar estimations can of course be made for the other five parts of the test, and an estimated total vocabulary size can subsequently be calculated. The accuracy of the estimate depends to a great extent on the number of items tested. In terms of total sample ratio, the 120-item test samples one word in 75 from an 8900-word vocabulary (the total number of words in Table 3).

Before we continue the description of the 120-item test, the sample ratio may be compared with the sample ratio of the widespread Vocabulary Levels Test (Nation 1983; 1990). The Levels test is a measure of vocabulary size. It has been called "the nearest thing we have to a standard test in vocabulary" (Meara 1996:38). The format consists of five parts, each relating to a particular frequency level of English. These levels are the first 2000, 3000, 5000, 10,000

words and a level called the university word level which is fitted in between the 5K and the 10K levels. For each level, 18 words are tested in a multiple-choice fashion. For the five levels together, 90 words are tested. This renders a sample ratio of one word in 110 from a 10,000-word vocabulary.

Returning to the 120-item test under investigation here, it measures learners' receptive knowledge of the target word with reference to what L1 word can be associated with it. Since the learner is provided with options, in the case at hand five of them, she doesn't have to *recall* the L1 translation equivalent from long-term memory, but rather *recognise* it. A recall task is generally considered to be more difficult than a recognition task, but the explanations given for why this is so are not conclusive (Nation 2001: 28). Since no context is given, testees need a more precise understanding of the target word than if contextual clues are provided, where strategies like meaning inferencing may play a role.

If the learner does not know the correct answer, despite the aid she gets from the alternatives, she might resort to guessing. As was pointed out in the previous subsection, the guessing factor is one of the identified problem areas of the multiple-choice format. Nation (ibid.: 349-350) reports of research carried out on L1 learners concerning answer strategies during multiple-choice tests. The research, which compared high-ability (ha) and low-ability (la) readers, showed that 'knowing the answer' accounted for 8 (la) and 16 (ha) per cent of the items, 'guessing the answer' accounted for 21 (la (50 per cent success rate)) and 8 (ha (35 per cent success rate)) per cent of the items. The conclusions drawn were that guessing is not a major problem and that some sort of knowledge is the driving factor behind learners' responses. Similarly, Mobärg (1997: 215) provides a couple of hypothetical calculation examples aimed at taking the sting out of the most critical remarks about guessing being a decisive factor. In one of his examples, Mobärg estimates that the chance of someone achieving 60 points out of the total 120 on the Gothenburg test, by guessing her way through the whole test, is one in a million. However, Mobärg's calculations are open to a minor objection. The fact that a test-taker may eliminate some of the distractors in an item, thus increasing her chances of making a successful guess, sometimes quite radically, is not taken into account in Mobärg's examples.

Having described and discussed the basic properties of the commonly used test, it is now time to take a look at how learners perform on it, and, by doing that, investigate aspects of validity of the test.

5 Learners' performance and test validity – the case of a test given in Lund in December 2003

This section reports an analysis of the scores of 90 learners on the 120-item multiple-choice vocabulary test described in subsection 4.2 above. An account of the test in terms of item content will be given, and the learners' scores will be

analysed and compared in terms of learner group affiliation. Furthermore, a check for implicational scaling will be performed.

The version of the test investigated here was given in December 2003 at the English department at Lund University and a total of 235 students took the test as an obligatory part of their studies of English on levels A, B and C. These levels are equivalent to the first, second and third term of fulltime studies, respectively, of English. Each term lasts for 4.5 months. The test is administered at the end of each study term and a re-sit is offered just before the start of a new term. The test in December was the regular, obligatory end-of-term test for all students of English on levels A-C[8]. In terms of proficiency, learners are expected to perform better for each term of study. Thus, one would expect students at level C to be more advanced than those at level B, who in turn would be expected to be more advanced than those at level A.

In total, the scores of 90 students were investigated, 30 students randomly selected from each of the three levels of study. The number was governed by the total number of students at the C-level taking the test. The student groups will henceforth be referred to as group A, group B and group C. The total score of each learner who took the test was readily available, but the score sheets from the selected 90 students were retrieved and marked a second time to eliminate any scoring errors by the person initially marking the test.

The maximum score on the test is 120. A score of 1 is given for each item answered correctly and 0 for each item answered incorrectly or not answered at all. If two choices are picked as response to the target word, then a score of 0 is given.

Since the test is given several times each academic year, the target words and accompanying distractors are different from test date to test date, but the samples of words for each part of the test are consistent from the designated frequency bands for that part. The target words of the particular version of the test investigated belonged to four parts of speech, all of them content words: nouns (69 = 58%), verbs (35 = 29%), adjectives (14 = 11%) and adverbs (2 = 2%).

In Table 4 below, descriptive statistics based on the 90 learner scores is presented.

As can be seen in the table, arithmetically, the total mean for group C was higher than for groups A and B, whose total mean scores were very similar, a few decimal points' advantage for group A over B, but with a greater standard deviation noted for group A.

In order to go beyond 'eye-ball statistics', however, and find out if the differences between the groups were truly significant, a One-way ANOVA

[8] This statement needs qualifying. In reference to Table 7 below, a student who scores a high pass on one level may transfer that result to automatically pass the next higher level, without actually having to do that test.

analysis was conducted. The null hypothesis (H_0) we want to reject is that there is no difference in vocabulary scores for different learner groups. The results from the ANOVA are given in Table 5 below:

Part		**1**	**2**	**3**	**4**	**5**	**6**	**Tot.**
Number of items		20	20	20	20	20	20	120
Group A	**Mean**	15.7	14.6	13.8	9.8	10.9	4.7	69.5
(n=30)	**S.d.**	2.6	2.4	2.8	4.3	2.9	3.3	15.8
Group B	**Mean**	16.1	14.2	13.4	9.9	10.4	5.2	69.2
(n=30)	**S.d.**	2.2	3.0	2.1	3.9	2.6	2.9	12.7
Group C	**Mean**	16.8	16.1	16.4	12.8	11.9	6.8	80.8
(n=30)	**S.d.**	2.3	2.0	2.0	4.0	2.9	3.4	12.9

Table 4. Results of the 120-item multiple-choice vocabulary test.

Source of variance	**SS**	**df**	**MS**	**F**	**Significance**
Between groups	2646.667	2	1323.333	6.879	.002
Within groups	16735.833	87	192.366		
Total	136.16	89			

Table 5. One-way ANOVA analysis of vocabulary scores across 3 learner groups

Table 5 tells us that we can reject the H_0. There is a significant difference between the group means ($p < .05$). However, we cannot know for certain where the differences lie. In order to find this out, a post hoc test (Tukey) was computed. The results of this test are given in Table 6. As the table shows, significant differences were found between groups A and C, and between groups B and C, but not between groups A and B, at the level of $p < .05$.

Group	**Group A**	**Group B**	**Group C**	**Significance**
A		.3333		.995
B			11.6667	.005
C	11.3333			.006

Table 6. Tukey test of differences across learner groups.

The students' results on the test and the comparison of the three groups call for a short discussion. The learners from the most advanced level of study,

group C, clearly performed better than the two less advanced groups. However, there was no significant difference in performance between groups A and B. Thus, the investigated version of the test did not differentiate between students at the less advanced learner levels (A and B). There can be several explanations to why this is so.

Firstly, only one term of study separates the groups from each other, and a clearly detectable increase in the type of vocabulary knowledge tested may not occur in such short period of time. However, the way the test is administered, at the end of each term of study, implies that an increase is expected. In terms of pass marks for the test, the following cut-off numbers were used:

	Fail	Pass	High Pass
Study level A	<62	≥62	≥80
Study level B	<72	≥72	≥90
Study level C	<82	≥82	≥100

Table 7. Cut-off numbers for different marks

In relation to the observed similarities between the mean scores for groups A and B, we can deduct from Table 7 that learners are required to score increasingly higher for each level of study in order to pass the test. Thus, for a pass, a student in group B needs a score which is ten points higher than that of a student in group A. As was mentioned in subsection 4.2, the test samples one word in 75. Thus, hypothetically, knowing ten more words in the test implies knowing about 750 more words from the total 'list', from which the sample items are taken (see subsection 2.1 for the assumption behind this estimate). Admittedly, estimates like these are extremely coarse and one should be careful not to read too much into them. On each level of study, students take courses in literature and text analysis, requiring them to read about 2000 pages of modern fiction and ordinary prose. In these courses, students are encouraged to work actively with vocabulary as a preparation for the upcoming vocabulary test. Consequently, a lack of significant difference between the groups can perhaps be seen as somewhat disappointing from a curriculum administrator's perspective. It should be emphasized, though, acknowledging the small number of subjects in this study, it can at best serve as an incentive for larger-scale studies into the matter.

Secondly, we note that the standard deviation is higher for group A than group B. By taking a closer look at the individual scores of the learners in group A, we find that two of the 30 students scored very high compared to the others in the group (scores of 108 and 113). These scores clearly caused an increase in the mean score.

It is possible to compare the learners' scores in the present study with learners' scores from the study mentioned at the beginning of subsection 4.2, Zettersten (1979), which was a cross-national study of the vocabulary knowledge of first-year students of English in the Scandinavian countries. Such

a comparison is interesting since the same test was used as that under investigation here, although with a different sample of target items, but from the same frequency bands. Zettersten also reports of two other surveys: scores from a Swedish national survey of language proficiency carried out universities in the early 1970s, and a study of first-year students of English at Gothenburg University from 1971. Table 8 below shows the mean scores from the present study juxtaposed with the results reported by Zettersten. Mean scores for all the levels of the present study are given for sake of comparison. If available, the number of subject scores from which the mean was calculated is given within parentheses after the mean scores.

Level of study	Present study (Dec 2003)	Zettersten study (Oct 1976)	National Survey study (Sep 1973)	Gothenburg study (1971)
Level A	69.5 (n = 30)	69.7 (n = 57)	67.0	63.9 (n = 240)
Level B	69.2 (n = 30)	-	-	-
Level C	80.8 (n = 30)	-	-	-

Table 8. A comparison of mean scores across different studies

The comparison of the mean scores suggests a pattern where students at level A seem to perform gradually better over time. However, an explanation for the somewhat differing mean scores may be that different versions of the test were employed in the different studies, a fact which makes straightforward comparisons rather doubtful. It is for example possible that the test used for the 1971 study might have been slightly more difficult than the one in the present study. Also, the rather small number of subjects restricts any firm conclusions to be drawn.

5.1 Implicational scaling

The validity of a test hinges on the extent to which it genuinely measures what it is set out to measure. Validation of tests can be done in many ways. One way of validating a test which samples words from different frequency bands is to analyse test-takers scores on the different parts of the test and investigate the score patterns. Since the underlying assumption of the test is that low-frequency words will be more difficult for learners than high-frequency words, we would expect learners to perform better on part 1 than on part 2, and better on part 2 than part 3 etc. Ideally, learners' scores will form a so-called implicational scale, which means that anyone getting a high score on for example part 4 of the test, is expected to have received equally high or higher scores on the preceding parts (1-3).

In order to check the level of implicational scaling, a Guttman procedure analysis was carried out (see Hatch and Lazaraton 1991: 204-212). First, a criterion score of 17 out of 20 for each level was set. This was done as the

Guttman analysis implies finding a scale in a set of dichotomous items. For the purpose of our analysis the scores were converted into values of either 1 or 0. A student who scored at least 17 on any part of the test was assumed to know practically all the words in the frequency band from which the sample was drawn (cf. Read 2000: 121-122), and her score was subsequently given a value of 1. In all cases where the criterion score was not achieved a value of 0 was given. Firstly, in any case where the criterion score was met on a part of the test which followed a part where it was not met, an error was noted. The analysis resulted in 19 errors, that is, nineteen breeches of the implicational pattern. But the number of errors observed will not tell us what we want to know, since several other factors need to be taken into account, e.g. number of subjects and number of items. Secondly, therefore, statistical computations were carried out. The Guttman procedure involves a series of calculations which eventually determine the *scalability coefficient*, which is the figure telling us if our data are scalable. Conventionally, this coefficient must be above .60 before scalability can be claimed. The analysis of the students' scores showed a coefficient of .78, which indicates that the data are scalable, even highly scalable. Thus, we may conclude that in this respect the test works along the lines of the underlying assumption upon which it is based.

To sum up, the analysis of the students' performances on the 120-item multiple-choice test has rendered two basic findings. Firstly, a comparison of the test scores of the three learner groups showed that learners on the most advanced level of study (group C), the 3rd term students, performed significantly better than learners on less advanced levels (groups A and B), and that little to no difference was found between 1st term students (group A) and 2nd term students (group B). Secondly, a validation analysis focusing on implicational scaling showed that the students' scores were highly scalable, with a scalability coefficient of .78. This tells us that the test in this respect works according the underlying assumptions of increasing difficulty for each 20-item part, due to decreasing level of frequency of the words tested.

6 Concluding remarks

The aim of this paper was to investigate whether reports in the literature about trends in the way L2 vocabulary is tested holds true for courses of English as a L2 at Swedish universities today. Leading scholars in the field of L2 vocabulary testing trace a move away from discrete vocabulary testing towards more embedded and comprehensive methods.

The investigation showed that the great majority, 19 out of 25, of the Swedish university organizers of English courses: English departments, modern languages departments and the like, use discrete vocabulary test methods. This thus goes against the claim in the L2 vocabulary testing literature. The investigation also showed that the multiple-choice format was the most

commonly used discrete test format. In particular, a specific frequency-based 120-item multiple-choice test, developed by researchers at Gothenburg University in the late 1960s, was in extensive use at Swedish universities. No less than 10 departments were found to use this specific test as a measure of L2 English vocabulary. The format and design of this test was investigated and accounted for, together with its underlying assumptions.

An analysis of test scores from university learners on three different levels of study showed that the test differentiated significantly between the 3rd term level and the 1st and 2nd term levels, whereas no significant difference could be observed between the 2nd term and 1st term level learners. The small number of subjects in the study restricts the consequences of the result but it may serve as a point of departure for further research. There seems to be a need for further studies of learner scores on the test, where learners at different levels of study are investigated. Possible reasons for why, for example, no difference between 1st term and 2nd term students could be established may be followed up in terms of studies involving more subjects. Analyses of what items may be known better at one level of study than at another, for example items belonging to different parts of speech, would render a better picture of learners taking the test. Also, another priority could be to control for external factors that may influence the results.

A Guttman procedure analysis for implicational scaling resulted in a high scalability coefficient that gave validation support to the underlying assumption of the test design. The assumption involves an expected correspondence between ascending order of difficulty and descending order of word frequency, and in this respect, based on the present small-scale analysis, the test works well.

7 References

Anderson, and P. Freebody. 1981. Vocabulary Knowledge. In *Comprehension and Teaching: Research Reviews*, ed. J. T. Guthrie, 77-117. Newark, DE: International Reading Association.

Chapelle, C. 1998. Construct definition and validity inquiry in SLA research. In *Interfaces Between Second Language Acquisition and Language Testing Research,* ed. L. Bachman & A. Cohen. 32-70. Cambridge: Cambridge University Press.

Coxhead, A. 1998. *An Academic Word List*. Occasional Publication Number 18. LALS. Victoria University of Wellington, New Zealand.

Coxhead, A. 2000. A new academic word list. *TESOL Quarterly* 34, 213-239.

Cronbach, L.J. 1942. An analysis of techniques for diagnostic vocabulary testing. *Journal of Educational Research* 36: 206-217.

D'Anna, C. A., E. B. Zechmeister, J. W. and Hall. 1991. Toward a meaningful definition of vocabulary size. *Journal of Reading Behaviour: A Journal of Literacy* 23: 109-122.

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

Ellegård, A. 1960. Estimating Vocabulary Size. *Word* 16: 219-244.

Francis, W.N. and H. Kucera. 1982. *Frequency analysis of English usage*. Boston, MA: Houghton Mifflin.

Goulden, R, P. Nation and J. Read. 1990. How large can a receptive vocabulary be? *Applied Linguistics* 11: 341-363.

Hatch, E. and A. Lazaraton. 1991. *The Research Manual – Design and Statistics for Applied Linguistics*. Boston: Heinle and Heinle.

Meara, P. 1990. A note on passive vocabulary. *Second Language Research* 6: 150-154.

Meara, P. 1996. The dimensions of lexical competence. In *Performance and Competence in Second Language Acquisition*, ed. G. Brown, K. Malmkjaer and J. Williams, 35-53. Cambridge: Cambridge University Press.

Meara, P. and Wolter, B. In Press. V_Links: beyond vocabulary depth.

Melka, F. 1997. Receptive vs. productive aspects of vocabulary. In *Vocabulary: Description, Acquisition and Pedagogy*, ed. N. Schmitt and M. McCarthy, 84-102. Cambridge: Cambridge University Press.

Mobärg, M. 1997. Acquiring, teaching and testing vocabulary. *International Journal of Applied Linguistics*. 7:2: 201-222.

Nation, I.S.P. 1983. Testing and teaching vocabulary. *Guidelines* 5: 12-15.

Nation, I.S.P. 1990. *Teaching and Learning Vocabulary*. New York: Heinle and Heinle.

Nation, I.S.P. 1993. Using dictionaries to estimate vocabulary size: essential, but rarely followed procedures. *Language Testing* 10:1: 27-40.

Nation, I.S.P. 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Read, J. 1997. Vocabulary and testing. In *Vocabulary: Description, Acquisition and Pedagogy*, ed. N. Schmitt and M. McCarthy, 303-320. Cambridge: Cambridge University Press.

Read, J. 2000. *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Richards, J. 1976. The role of vocabulary testing. *TESOL Quarterly* 10: 77-89.

Schmitt, N. 2000. *Vocabulary in Language Teaching*. Cambridge: Cambridge University Press

Thorén, B. 1967. *10 000 ord för tio års engelska*. Lund: Gleerups.

Thorndike, E. and I. Lorge. 1944. *The Teacher's Word Book of 30,000 Words*. New York: Teachers College, Columbia University.

Waring, R. 1997. A comparison of the receptive and the productive vocabulary sizes of some second language learners. *Immaculata* 1: 53-68.

Waring, R. 1999. *Tasks for Assessing Second Language Receptive and Productive Vocabulary*. Unpublished Doctoral Thesis: University of Wales, Swansea.

West, M. 1953. *A General Service List of English Words*. London: Longman.

Zettersten, A. 1979. *Experiments in English Vocabulary Testing*. Malmö: LiberHermods.