

Testing Vocabulary Knowledge: Size, Strength, and Computer Adaptiveness

Batia Laufer and Zahava Goldstein
University of Haifa

In this article, we describe the development and trial of a bilingual computerized test of vocabulary size, the number of words the learner knows, and strength, a combination of four aspects of knowledge of meaning that are assumed to constitute a hierarchy of difficulty: passive recognition (easiest), active recognition, passive recall, and active recall (hardest). The participants were 435 learners of English as a second language. We investigated whether the above hierarchy was valid and which strength modality correlated best with classroom language performance. Results showed that the hypothesized hierarchy was present at all word frequency levels, that passive recall was the best predictor of classroom language performance, and that growth in vocabulary knowledge was different for the different strength modalities.

Vocabulary Knowledge and Vocabulary Tests

Vocabulary tests are contingent upon the test designer's definition of lexical knowledge. Lexical knowledge, in turn, has

Batia Laufer, Department of English Language and Literature; Zahava Goldstein, Department of Education.

The study presented in this article is a follow up to Laufer, Elder, Hill, and Congdon (2004), which researched the monolingual Computer Adaptive Test of Size and Strength.

Correspondence concerning this article should be addressed to Batia Laufer, Department of English Language and Literature, University of Haifa, Haifa 31905, Israel. Internet: batialau@research.haifa.ac.il

been defined differently by different researchers. It has often been defined as the sum of interrelated “subknowledges”—knowledge of the spoken and written form, morphological knowledge, knowledge of word meaning, collocational and grammatical knowledge, connotative and associational knowledge, and the knowledge of social or other constraints to be observed in the use of a word (Nation, 1990, 2001; Richards, 1976; Ringbom, 1987).

Alternatively, lexical knowledge has been construed as a continuum consisting of several levels of knowledge, starting with superficial familiarity with the word and ending with the ability to use the word correctly in free production (Faerch, Haastrup, & Phillipson, 1984; Palmberg, 1987). One researcher, Henriksen (1999), perceives lexical knowledge in terms of three continua: partial to precise knowledge, shallow to deep knowledge, and receptive to productive knowledge. She does not consider word use to be a kind of knowledge, but rather to be control over word access.

Most vocabulary tests that are based on the knowledge components model measure just one of the subknowledges: comprehension of meaning (Meara & Buxton, 1987; Nation, 1983), production of meaning (Laufer & Nation, 1999), vocabulary use (Arnaud, 1992; Laufer & Nation, 1995), or word associations (Read, 1993). Some tests, however, attempt to measure simultaneously several subknowledges (Read, 1989; Schmitt, 1999). Few tests attempt to trace the learner’s progress along a continuum of knowledge (but see Schmitt, 1998; Wesche & Paribakht, 1996).

When just one component of knowledge is tested, it is usually possible to test a large sample of items, and the test can thus claim to represent the learner’s total vocabulary. Such tests are called vocabulary “breadth” or “size” tests. However, these tests have been criticized for the superficial treatment of each item, that is, testing an item on one component of knowledge only. The alternative to a size test is a “depth” test, on which each item is tested on several components of knowledge.

The limitation of depth tests is that the number of items that can be tested is limited and the test does not, therefore, represent the true vocabulary of the test taker.

Meara (1996) and Laufer and Nation (2001) have suggested that an additional component of word knowledge (referred to as a “dimension” in these studies) is automaticity of access, or fluency, or the speed with which a person can perform some kind of operation on a word. Consequently, they have designed tests that try to measure fluency of a large number of words.

All of the studies mentioned above approach vocabulary knowledge as the knowledge of discrete word items independent of context in which they appear. Such a view of vocabulary is referred to by Chapelle (1998) as the “trait” view. Contrary to the trait view, Read (1997, 2000) and Read and Chapelle (2001) propose to approach vocabulary assessment from an interactionist perspective. Following Bachman’s (1990) view of language proficiency as a set of communicative skills, they suggest that the lexical skill should incorporate communicative competence in addition to the knowledge of discrete items. They claim that a vocabulary test should be defined in relation to a particular context typical of the test takers’ needs, that it should go beyond knowledge of decontextualized word lists, and that the test should have positive consequences, such as giving learners “the incentive to deepen their knowledge of lexical items, or to develop effective communication strategies” (Read & Chapelle, 2001, p. 23).

To summarize, different researchers recommend different vocabulary tests, depending on their view of vocabulary knowledge (trait or interactionist), their preference for a particular dimension of knowledge, and their interest in either size or depth. In this article, we adhere to the trait view of vocabulary, and though we acknowledge the importance of depth and fluency tests, we choose to focus on vocabulary size. Tests of vocabulary size have been shown to predict success in reading, writing, and general language proficiency as well as academic achievement (Laufer, 1997; Saville-Troike, 1984), whereas other types of

vocabulary tests as yet have not. Therefore, size tests allow for more efficient placement and admission in language teaching programs than depth tests or fluency tests. Moreover, size tests can function as relatively simple and efficient research instruments providing us with the vocabulary size of the participants at the beginning of the treatment and showing the growth in vocabulary after an experimental intervention. Even though Read (2000) has worked on developing depth tests, he points out that “despite the fact that the [size] tests may seem superficial, they can give a more representative picture of the overall state of the learner’s vocabulary than in-depth probe of a limited number of words” (p. 115).

Form-Meaning Link in Vocabulary Tests

All the researchers surveyed above realize that word knowledge involves more than just the link between meaning and form. In spite of this, most vocabulary tests focus, directly or indirectly, on this link. Direct tests of the form-meaning link are tests in which the learners are required to demonstrate their understanding of the target words or produce the target form for given meanings (see, for example, Laufer & Nation, 1999; Meara & Buxton, 1987; Nation, 1983; Schmitt, Schmitt, & Clapham, 2001). In such tests, the target word is matched with its paraphrase, translated into the test taker’s first language (L1), marked *yes* if understood and *no* if not, or supplied in a sentence with first letters provided.

More interesting are indirect tests of the form-meaning link. These do not look like word-meaning tests (Arnaud, 1992; Laufer & Nation, 1995; Read, 1993; Schmitt, 1999; Wesche & Paribakht, 1996), yet the form-meaning link is central to what is being tested. For example, a test of associations, whether alone or in conjunction with other areas of knowledge, also tests whether the meaning of the target word is known. In such a test (Read, 1993), the learner is asked whether the word tested is related to any of a group of eight other words, for example,

whether *edit* is associated with *article* and *publish*. The correct decision can be made only if the meaning of *edit* is known. Wesche and Paribakht (1996) claim that the Vocabulary Knowledge Scale is a test of vocabulary knowledge depth. We argue that it is an indirect test of word meaning. One category of the scale states, "I have seen the word but I can't remember what it means"; another, "I think I know what it means"; and yet another, "I'm sure I know what it means." So far, the test has tested the word's meaning. The most advanced stage of the scale requires the learner to write a sentence using the target word. The authors' assumption is that sentence writing demonstrates several areas of word knowledge, in addition to understanding of the word's meaning. However, the problem with sentence writing is that, in many cases, a sentence reveals not much more than knowledge of meaning. For example, the unique grammatical feature of *news* being a singular is not revealed in the sentence *Every morning I listen to the news*. Whether a test taker is familiar with the collocation *heavy traffic* is not clear from the sentence *I hate traffic in the morning*. All that such sentences show about the knowledge of *news* and *traffic* is that the student understands the referential meaning of these words rather than their various semantic and grammatical frames. Laufer and Nation's (1995) Lexical Frequency Profile (LFP) is also an indirect test of meaning. The authors argue that the proportion of frequent to infrequent vocabulary in a text reflects the proportion of frequent to infrequent vocabulary in the lexicon of the text producer. However, because only correct vocabulary is included in the profile (words with semantic errors are not), the LFP shows the proportion of frequent versus infrequent correct form-meaning links.

The emphasis of vocabulary tests on meaning is inevitable, whether researchers acknowledge it or not, because words are the first and foremost units of meaning. A student who knows what *advice* means, but does not know that it is used as an uncountable noun, and says, **"The mother gave her daughter many advices,"* will be understood in spite of the

grammatical error. On the other hand, a student who knows that *advice* is used in singular but confuses its meaning with *advance*, for example, will experience a break in communication. The knowledge of word meaning is, therefore, the central component of word knowledge, and a good vocabulary test should measure the extent to which those who take it can correctly associate word form with the concept the form denotes.

Strength of Knowledge of Meaning

Most vocabulary researchers distinguish between passive (receptive) and active (productive) knowledge of a word (e.g., Meara, 1990; Nation, 2001). On the surface, the distinction looks simple. Passive knowledge is associated with listening and reading and implies that we are able to comprehend the input. In vocabulary, this means that we are able to perceive the form of the word and retrieve its meaning or meanings. Active knowledge is associated with speaking and writing and implies that we can retrieve the appropriate spoken or written word form of the meaning that we want to express. Though listening and reading may not be entirely “passive” processes, the terms *passive* and *receptive* are often used interchangeably and will be so used in this article. An example of testing passive word knowledge is asking for a word’s L1 equivalent, and an example of testing active knowledge is asking for the second language (L2) equivalent of a given L1 word (Takala, 1984). In multilingual classes, or when the tester is not familiar with the students’ L1, an L2 paraphrase may be asked for to demonstrate passive knowledge, and an L2 target word for a synonym or a paraphrase may be required to prove active knowledge. Another common testing method is asking the learner to select the correct answer from several meaning options for the given word or to select the correct answer from several word forms for a given meaning (e.g., Waring, 1997). Whereas Takala’s tests are recall

tests, Waring's are recognition tests. Another test of active knowledge is asking for an original sentence in which a given word is to be used. Being able to produce a sentence with the correct use of the word is sometimes viewed as the most advanced state of word knowledge.

Yet the distinction between passive and active knowledge is not as simple as it may appear. Read (2000) points out that different people have made this distinction differently, which has created problems in comparing passive and active knowledge across studies. For example, sometimes translating into L1 is considered active knowledge, since supplying the meaning requires the test taker to produce something. In such cases, test results are sometimes reported as active-knowledge results (Schmitt, 1999). At other times, the same test is considered a test of passive knowledge, since the test takers demonstrate comprehension of the meaning, rather than availability of the L2 word form (Takala, 1984).

Another problem with the active-passive knowledge distinction is that there is no consensus as to whether this distinction is dichotomous or whether it constitutes a continuum. Melka Teichroew (1982) argues that the distinction is arbitrary and proposes to treat passive and active vocabulary as poles on a continuum representing increasing degrees of word knowledge. Yet it is not clear how much knowledge is necessary for a word to move from passive to active status. Meara (1990), on the other hand, regards the two types of knowledge as representing fundamentally different kinds of associational knowledge: Active vocabulary is activated by other words in the language, whereas passive vocabulary can be activated by external stimuli only.

In this article, we will try to overcome the confusion between active and passive vocabulary and will distinguish among four degrees of knowledge of meaning. We will also hypothesize and show that they constitute a hierarchy. The four degrees of knowledge are based on two dichotomous distinctions:

- supplying the *form* for a given meaning versus supplying the *meaning* for a given form

- being able to *recall* versus only being able to *recognize* (whether form or meaning)

The first distinction implies that there is a difference in knowledge between people who can supply an L2 word and those who can only supply the meaning when the L2 word is presented to them. We will refer to the ability to supply the word form as active knowledge and to the ability to supply the word meaning as passive knowledge.¹ The second distinction implies that there is a difference in knowledge between those who can recall the form or the meaning of a word and those who cannot recall but can recognize the form or the meaning in a set of options.

The two distinctions mentioned above allow us to differentiate four degrees of knowledge of meaning, as illustrated in Table 1. The following subsections provide an example of how an item can be tested on the four degrees of knowledge in a bilingual test. The target item is *affluence*.

Active recall. The task in active recall is to supply the L2 target word. The prompt is the L1 translation equivalent. In our example, the L1 is Hebrew. The first letter of the target word is provided to prevent the student from supplying nontarget words that have the same meaning:

a_____ אפע

Passive recall. The task in passive recall is to demonstrate the understanding of the meaning of the L2 word. The L2 word is in the prompt, and the learner has to supply the L1 translation. The first letter of the translation is provided:

Affluence_____ א

Active recognition. The task in active recognition is to choose the target word from among four options. The prompt is the L1 translation of the target word. The options include the L2 target word and three distractors that are semantically unrelated to one another, since we do not intend to test comprehension of fine shades of meaning. The distractors are taken

Table 1

Degrees of vocabulary knowledge

	Recall	Recognition
Active (retrieval of form)	Supply the L2 word	Select the L2 word
Passive (retrieval of meaning)	Supply the L1 word	Select the L1 word

from the frequency level of the target word, which makes them as difficult for the learner as the target word:

שפע

a. precision b. affluence c. axis d. episode

It may be claimed that since nothing is “produced” (said or written) by the learner, the term *production* or *active knowledge* may be inadequate. Yet productive or active knowledge is defined by Nation (2001, p. 359) as knowledge “that is used in speaking and writing, and involves going from the meaning to the word form.” Nation argues that productive knowledge can be checked using a recall task in which the target word has to be recalled, as in translating a word from L1 into L2, or using a recognition task in which the target word has to be recognized and selected from among several options. We therefore consider this type of recognition as active.

Passive recognition. In passive recognition, the target word is given as a prompt, and the task is to choose its meaning from among four options in the learner’s L1. The correct response is the Hebrew translation of *affluence*. The distractors are L1 translations of the distractors that appear in the productive recognition task (axis, episode, precision):

affluence

שפע ציר פרק דיוק

The idea of strength of knowledge tries to overcome the limitation of vocabulary size tests in which words are tested in one modality only, usually passive recognition. We hypothesize that knowledge of meaning constitutes a hierarchy in which

some degrees of knowledge are more advanced than others and presuppose the less advanced degrees of knowledge.

L2 learners' passive vocabulary is larger than their active vocabulary (Laufer, 1998; Laufer & Paribakht, 1998). This means that many words are first acquired passively and that active knowledge is a more advanced degree of knowledge. A person who can retrieve the word *form* for a given concept is typically able to retrieve its *meaning* upon encountering the form. As for the recall/recognition dichotomy, recall of information indicates a better memory trace than recognition of the same information. Hence, recalling a word's meaning or form can be considered a more advanced degree of knowledge than recognizing it in a set of options. Language learners who can recall the meaning of a given word can typically recognize the meaning among several options (if the options are not intentionally designed to confuse the learner). The four degrees of knowledge will, from now on, be referred to as degrees of "strength" of knowledge. If active knowledge is more difficult to achieve than passive knowledge, and if recall is more difficult than recognition, then the most advanced degree of knowledge is reflected in active recall and the least advanced knowledge is passive recognition. As for the two remaining degrees of strength, passive recall and active recognition, intuitively, we felt that passive recall requires more knowledge than active recognition, but before the validation study we did not hypothesize which of these degrees of strength was more difficult to attain. The four degrees of strength of knowledge are presented in Table 2. Whether they form a hierarchy as shown in Table 2 is an empirical question that we address in this study.

Table 2

Four degrees of strength of knowledge

	Recall	Recognition
Active (retrieval of form)	(1) strongest	(2) or (3)
Passive (retrieval of meaning)	(2) or (3)	(4) weakest

Computer Adaptive Test of Size and Strength

Basic Assumptions

Three assumptions underlie the rationale for the test presented in this article:

1. The most important component of word knowledge is the ability to establish the *link between word form and word meaning*.
2. Knowledge of meaning is not an all-or-nothing phenomenon. The form-meaning link in the mental lexicon can have four degrees of strength: active recall (which is the strongest and presupposes the other three degrees of strength), passive recall, active recognition, and passive recognition, which is the weakest.
3. Knowing *many* words (units of meaning) is more important than knowing few words *in depth*. Hence, a good vocabulary test should test *how many words* are known. In other words, it should try to provide a picture of the learner's overall vocabulary.

The test we are proposing is a vocabulary size test of word meaning. However, unlike other size tests, which focus on one area of meaning, we test each word in more detail, thus overcoming the basic limitation of size tests. In the Computer Adaptive Test of Size and Strength (CATSS), each word is tested in four modalities, representing four degrees of strength of knowledge, as illustrated earlier using the word *affluence*. CATSS is therefore a test of vocabulary size and strength.²

Test Description

The main features of CATSS are vocabulary size, strength, and computer adaptiveness. The test was designed in two versions: monolingual and bilingual.

Vocabulary size. CATSS includes samples of items from five word frequency levels (30 items from each level): the 2,000 most frequent words, the 3rd thousand, the 5th thousand, the 10th thousand, and the Academic Word List (AWL), which cuts across words from the 3rd and 5th thousands and includes some less frequent words as well. These are the levels established by Nation in his Vocabulary Levels Test (1983), which is widely used for teaching and research purposes. Nation's Vocabulary Levels Test has 18 items at each level. We, however, used 30 items at each level, following the findings of Schmitt et al. (2001) in regard to their recently validated version of the Vocabulary Levels Test. In their version of the test, the items were randomly selected from frequency lists. In the process of validation, it was found that 30 items were sufficient to be considered representative of each level. Since CATSS includes five levels of vocabulary and 30 items at each level, in total, 150 words are tested. Since each word is tested in four modalities (active recall, passive recall, active recognition, passive recognition), there are 600 items on the entire test. The test for the 2,000-word level is shown in Appendix A.

Vocabulary strength. The vocabulary *strength* element is reflected in four test stages, each test stage testing one modality for each word, starting with the most difficult one (active recall) and ending with the easiest one (passive recognition). Table 3 presents the structure of CATSS.

Table 3

Structure of CATSS

Active recall	Passive recall	Active recognition	Passive recognition
2,000	2,000	2,000	2,000
3,000	3,000	3,000	3,000
5,000	5,000	5,000	5,000
10,000	10,000	10,000	10,000
Academic	Academic	Academic	Academic
Word List	Word List	Word List	Word List

Computer adaptiveness. CATSS has been designed as a computerized test. To validate the hierarchy of strength of knowledge, however, the test had to be administered on paper, since each student had to be tested on all four strength modalities of a word, and in the computerized version, a correct answer on one modality means that the item is not tested on the subsequent modalities. For example, if a test taker has correctly recalled a particular target word in active-recall mode, that word will not appear again in the other three modalities (passive recall, active recognition, and passive recognition). If, however, no response or an incorrect response is given for a particular word in the active-recall section, the computer keeps the word in its memory for presentation in the next modality. The word is not presented again immediately after the lacking or incorrect response. It appears on the screen only after all the other words of the same frequency level have been tested in the same modality. Once an item involving the word has been answered correctly in one of the modalities, the word is not presented again in subsequent modalities. After the four modalities of one frequency level have been tested, the next frequency level appears on the screen in its four modalities. Table 4 presents the four cycles of testing at each level of frequency.

The aim of the study was to validate the assumed hierarchy of strengths of knowledge for the four cycles of testing as

Table 4

Adaptiveness of CATSS to strength of knowledge

Strength modality	Correct items	Incorrect items
(1) Active recall	No further testing	Presented again at modality (2)
(2) Passive recall	No further testing	Presented again at modality (3)
(3) Active recognition	No further testing	Presented again at modality (4)
(4) Passive recognition	No further testing	No further testing

presented in Table 4. Therefore, for each word, an answer in each mode was elicited from the participants. This could be done on paper only, as the computerized version would stop testing a word once it was correctly answered in one of the modalities. If we found that the four strengths of knowledge were implicationally scaled, then we would be correct in our claim that the final computerized version would not need to test the same word in all four modalities.

In addition to the algorithm determining the cyclic testing of the four strengths of knowledge, the computer program for CATSS also has “end of test” and “jump level” functions. If learners do not do well on a low-level vocabulary, there is no point in presenting them with too many more difficult items. Therefore, the tester can decide on a cutoff point at which the test will end. Conversely, if a learner performs very well on a low level (i.e., on high-frequency words), there is no need to test the entire sample of words at that level. The tester can decide where the passage to the next level will occur. In this study, we do not deal with these functions of CATSS but leave them for further research.³

Scoring and displaying results. The answers provided by the test taker are matched against a prespecified marking key and are scored as correct or incorrect. In the bilingual CATSS, each item has one acceptable response. In active recall, the correct answer is the translation equivalent of the L1 prompt (the first letter of the answer is provided). In passive recall, the correct answer is the L1 translation of the L2 prompt (the first letter is provided). The prompt and the first letter restrict the given responses to the target items. A spelling error makes the answer incorrect. In the two recognition modes, one correct answer is selected from among four options.

As shown in Table 5, for each word frequency level, the learner receives four separate scores (one score for each strength modality). At the end of the computerized test, the program displays one to five such tables, depending on the number of frequency levels the student was tested on.

Table 5

Sample student result, Level 5,000

Strength modality	Number of correct responses (out of 30)
(1) Active recall	5
(2) Passive recall	22
(3) Active recognition	0
(4) Passive recognition	3

Table 5, which presents real data from the computerized version of CATSS, shows that out of the 30 items tested, all 30 were known in one way or another. In a conventional Vocabulary Levels Test, which is a test of passive recognition (Nation, 1983; Schmitt et al., 2001), the learner in Table 5 would have received a full score for this level. The advantage of CATSS is that it shows that not all the 30 items are known in the same way. Table 5 shows that the learner has a good vocabulary for reading, since he or she can recall the meaning of 27 out of 30 items at the 5,000-word frequency (5 items can be recalled for form, and recall of form presupposes recall of meaning, and 22 items can be recalled for meaning). However, we cannot be as confident about the student's ability to function on productive tasks, since only 5 out of 30 items were retrieved with their forms at the active-recall modality. This demonstrates the value of the test for educational and research purposes. If the learner in Table 5 is admitted into a course for academic purposes in an English-speaking country, he should not necessarily be placed at an advanced level, as his active vocabulary is not very good. If, however, the same learner enrolls in a course that emphasizes reading in an English as a foreign language (EFL) context, he should be considered more advanced. As a research instrument, CATSS taps the incremental nature of vocabulary knowledge. In a longitudinal study, we may find that learners progress differently on different strengths of knowledge or differently on different words.

In summary, CATSS has the advantages of a vocabulary size test, but it provides a more realistic picture of how well learners know the meaning of the tested items. Furthermore, computer adaptiveness makes the test efficient, since students will not need to be tested on all the items in each strength modality and students at different levels of proficiency will be tested on a different number of items.

The Study

The purpose of this research was twofold: (1) to investigate the validity of the hierarchy assumed in the sequence of the four strength modalities in the bilingual version of CATSS and (2) to examine the connection between the different strength modalities and academic success.

Research Questions

Our main research questions were as follows:

1. Is there a significant difference between the learners' scores on the four strength modalities? In other words, does active recall of a particular vocabulary item in fact presuppose passive recall, active recognition, and passive recognition of that item?
2. Can we assume that the hierarchy in the sequence of the four strength modalities is present at all word frequency levels?
3. Which of the four strength modalities can best explain the variance in foreign language class grades? In other words, which type of word knowledge is most closely related to L2 achievement?

An affirmative answer to Questions 1 and 2 would validate CATSS in terms of the proposed progression through the various strength modalities. As the test is currently conceived, each test

taker starts at active recall (i.e., the first and allegedly hardest strength modality) at the 2,000-word level and then works through each of the remaining strength modalities at the 2,000-word level, before proceeding to the active-recall modality of the 3,000-word level, then to the other modalities at that level, and so on. The answer to Question 3 is of practical value to program designers, teachers, and learners. It will show which strength modality is best associated with success in school. Consequently, teachers and learners may decide to concentrate their efforts on this specific word-meaning link. The correlations will also show the extent to which vocabulary knowledge is a good predictor of language grade. If the correlations are high, administrators may consider incorporating a vocabulary component, or enlarging the existing one, in language placement and admission tests.

Method

Data Collection

The tests were administered in pencil-and-paper format to 435 high school and university L2 learners. There were 278 native speakers of Hebrew, 140 native speakers of Arabic, and 17 native speakers of Russian. The last group knew Hebrew well enough to take the English-Hebrew test.⁴ The L1 speakers of Arabic took an English-Arabic version of the test.

The high school students were taking English classes as part of the national curriculum in Grades 9–12; that is, they had studied English for six (Grade 9) to nine (Grade 12) years. The university students, who were also high school graduates, were enrolled in English language courses designed to prepare them for academic study. Since it was clearly impractical to administer all test items to all participants, the participants were sampled using a nonprobabilistic method. Individual participants completed one frequency level in the four strength modalities.

Ninth and 10th graders received the 2,000-word level; 11th and 12th graders received either the 3,000-word level, the 5,000-word level, or the Academic Word List level. The university students received the 5,000-word level. (We did not have learners who were advanced enough to do the 10,000-word level.) Hence 480 items were tested in this study. Table 6 presents the sampling results.

The participants were presented with all 30 items at the level to which they were assigned in all four strength modalities.⁵ To minimize any practice effect, the order of the items in each subtest (testing each modality) was different, and the testing was distributed over two class sessions, with two subtests per session. The first subtest was the active recall. On the completion of the subtest, all test sheets were collected. Then half of the learners received passive recall, and half active recognition. This procedure was adopted because, as mentioned earlier, we did not hypothesize which of these two strength modalities was more difficult than the other.⁶ On completion of this subtest, the test sheets were collected. In the next subtest, students who received passive recall earlier, on the second subtest, received active recognition, and those who received active recognition earlier received passive recall. On completion of the subtest, the test sheets were collected and the last subtest was

Table 6

Distribution of participants by school grade and test level

Participant	9th grade	10th grade	11th grade	12th grade	University students	Total
Level						
2,000 word	52	82				134
3,000 word			83	23		106
Academic Word List			22	60		82
5,000 word			19	50	44	113
Total	52	82	124	133	44	435

administered: the passive recognition. For each participant, we recorded the score at each strength modality, the order in which the passive-recall and active-recognition modalities were presented, and his or her mother tongue. For 155 students, we also obtained an independent measure of language knowledge: a class grade given by the teacher. This grade was a combination of several class tests, which were given the heaviest weight, and also homework and class participation.

Results

The first research question asked whether there were significant differences among the learners' scores on the four strength modalities. The second research question asked whether the hierarchy in the sequence of the four strength modalities was present in each word frequency level. Table 7 presents separate mean scores for each frequency level and significance results for differences among modalities, as calculated using repeated measures.⁷

Table 7 shows that our assumed hierarchy of the modalities is not dependent on vocabulary frequency. At every frequency level, the hierarchy was the same: active recall yielded the lowest scores, and then in ascending order, passive recall, active recognition, and passive recognition. The differences among the four modalities were highly significant at each word frequency level ($p < .001$). Differences between pairs of modalities were tested using Tukey's post hoc test. Significant differences were found between each of the six possible combinations of modality pairs at each word frequency level. Since active recall is the most difficult modality and passive recognition is the easiest, the largest difference in means is between these two. The smallest (though significant) difference in means occurred between active and passive recognition.

The large differences among the modalities imply that most individual items are implicationally scaled, that is, they follow the hypothesized hierarchy of difficulty: active recall, passive

Table 7

Hierarchy of four strength modalities at each word frequency level

Modality Level	Active recall	Passive recall	Active recognition	Passive recognition	Significance of difference
2,000 word <i>n</i> = 134	<i>M</i> = 6.51 <i>SD</i> = 5.26	<i>M</i> = 18.61 <i>SD</i> = 8.43	<i>M</i> = 26.53 <i>SD</i> = 5.20	<i>M</i> = 27.79 <i>SD</i> = 3.58	<i>F</i> (3, 131) = 955.60 <i>p</i> < 0.001 Effect size = 2.63
3,000 word <i>n</i> = 106	<i>M</i> = 5.35 <i>SD</i> = 3.33	<i>M</i> = 17.58 <i>SD</i> = 5.99	<i>M</i> = 24.93 <i>SD</i> = 5.19	<i>M</i> = 27.80 <i>SD</i> = 3.73	<i>F</i> (3, 103) = 998.16 <i>p</i> < 0.001 Effect size = 2.84
5,000 word <i>n</i> = 113	<i>M</i> = 2.50 <i>SD</i> = 2.88	<i>M</i> = 15.60 <i>SD</i> = 6.88	<i>M</i> = 23.50 <i>SD</i> = 5.98	<i>M</i> = 28.60 <i>SD</i> = 3.17	<i>F</i> (3, 110) = 2067.29 <i>p</i> < 0.001 Effect size = 3.16
Academic Word List <i>n</i> = 82	<i>M</i> = 3.18 <i>SD</i> = 3.04	<i>M</i> = 13.07 <i>SD</i> = 8.43	<i>M</i> = 20.50 <i>SD</i> = 4.92	<i>M</i> = 26.79 <i>SD</i> = 4.41	<i>F</i> (3, 79) = 833.25 <i>p</i> < 0.001 Effect size = 2.55

Note. Maximum score in each cell = 30.

recall, active recognition, passive recognition. We sought additional evidence for the hierarchy by calculating Guttman's coefficient of reproducibility for 16 randomly selected words. We selected four words from each of the four frequency levels that were tested and checked the tests of four intact classes, one class for each frequency level test. For each word, we entered the responses of individual students, marking a plus sign for a correct response and minus sign for an incorrect response (see Appendix B). A word's responses are perfectly scaled if there are no plus responses to the left of minus responses, that is, if learners do not answer correctly on the more difficult modalities if they answered incorrectly on the less difficult ones. Deviations from the assumed scale were circled on the word sheet. The coefficient of reproducibility, which we calculated, tells us to what extent students' responses are in accordance with the assumed hierarchy. Mathematicians have determined that the value of the coefficient of reproducibility should be over .90 before a scale can be considered valid (Hatch & Lazaraton, 1991). In Table 8, we list the 16 words examined, and for each word, we present the following data: the number of students sampled for the procedure, the number of deviations from a perfect scale across all students in the sample, and Guttman's coefficient of reproducibility. Table 8 shows that the number of deviations is very small and that the scales for all the 16 words are valid.

The results in Tables 7 and 8 show that we can give a positive answer to the first and second research questions, which addressed the difference between the learners' scores on the four strength modalities.

The third research question asked which type of word knowledge was most strongly related to L2 achievement. Table 9 presents the mean class grades (in percentage) of the participants. In order to investigate which of the four modality scores could explain the largest variance in class grade in English, we carried out a stepwise multiple-regression analysis, with the four modality scores as the predictor variables and the

Table 8

Guttman's coefficient of reproducibility by word and frequency level

Level	Academic Word List															
	2,000 word				3,000 word				5,000 word				List			
Number of students	27				30				31				31			
Word sampled	A R R A N G E	B R A V E	V I C T O R Y	S E C R E T	A T M O S P H E R E	A N N U A L	T R A D I T I O N	A W A R E	A D E Q U A T E	E F F I C I E N C Y	R E S E N T	C O M P L I M E N T	C O N F I G U R A T I O N	P A R T I C I P A T E	A B S O R B	T R A N S F O R M
Number of deviations from scale	6	0	1	0	0	0	1	1	2	0	3	0	4	3	5	4
Guttman's coefficient <i>R</i>	.94	1.0	.99	1.0	1.0	1.0	.99	.99	.98	1.0	.98	1.0	.97	.98	.96	.97

class grade in English as the criterion variable. Table 10 shows the matrix of correlation between English class grades and the four strength modalities.

From Table 10, we can see that each modality was significantly correlated with the class grade. The highest correlation is

Table 9

Class grades

Participants	11th grade	12th grade	University students
<i>N</i>	41	89	25
<i>M</i>	74.36%	77.65%	73.16%
<i>SD</i>	14.78%	11.94%	9.78%

with passive recall. Stepwise regression analysis revealed that the total explained variance of the model was 42.6%. The modality that most contributed to the variance was passive recall (40%). Passive recognition added a modest 2.6% to the explained variance. The table also shows that all the strength modalities correlate significantly with one another. This suggests that the scores reflect basically the same trait, that of vocabulary knowledge. Yet as we have seen earlier, the actual scores on the various modalities are very different, and the modalities should therefore be tested separately.

Discussion

The central argument of this article is that the main component of vocabulary knowledge, which is the knowledge of form-meaning relationship, can be construed as a hierarchy of four degrees of strength: active recall, passive recall, active recognition, and passive recognition. The study presented in this article lends strong support to this hypothesis. Learners' performance was significantly different on each one of the four modalities, and there were also significant differences between each pair of modalities. Active recall yielded the lowest scores, passive recognition the highest. Additional support for the hierarchy hypothesis was provided by the results in regard to

Table 10

Correlation between class grade and vocabulary knowledge

Variable	Class grade	Active recall	Active recognition	Passive recall
Active recall	0.40**			
Active recognition	0.50**	0.53**		
Passive recall	0.63**	0.65**	0.65**	
Passive recognition	0.49**	0.31**	0.63**	0.58**

** $p < 0.001$.

the study's second research question. These showed that the sequence of the four strength modalities was present at all word frequency levels.

Researchers agree that vocabulary acquisition is incremental in nature and that some components of word knowledge are acquired before others. For example, the basic meaning of a word is usually acquired before its collocations and associations. There is almost no research that has explicitly shown a possible sequence in which components of knowledge are acquired. In Schmitt's (1998) study, for example, learners acquired words' spelling before their derivational forms and meanings. Our study demonstrated the incremental nature of acquiring form-meaning link. We found that the most advanced knowledge of meaning, which is acquired last, is active recall, the ability to recall a word's form. However, if the learner is unable to recall word form, this does not mean he or she cannot recall its meaning, as recall of meaning is a less advanced component of knowledge that can be acquired before recall of form. And before reaching the stage of passive recall, the learner may be able to recognize the word's form, and before then, to recognize the word's meaning.

The strength hierarchy that was hypothesized in the construction of CATSS had previously been tested with the monolingual version of the test. The monolingual version makes use of synonyms, paraphrases, and definitions and includes a subtest of the 10th thousand-word list. (For a detailed description of the monolingual version and its validation, see Laufer, Elder, Hill, & Congdon, 2004).

The results of Laufer et al. (2004) distinguished three rather than four different modalities. In that study, active recall proved to be more difficult than passive recall, followed by active and passive recognition, respectively. However, the results for the two recognition modalities were not significantly different from one another. The authors suggested that picking the correct definition of a word may not necessarily be easier than choosing the word form that matches a given set of definitions.

In the present study, however, the results for the two recognition modalities were significantly different from one another. A possible explanation for the difference between the recognition results in the studies of the monolingual and the bilingual versions of CATSS is that, in the passive-recognition subtest, choosing the correct translation of a word (in the bilingual version) is easier than choosing the correct definition (in the monolingual version). Therefore, in the bilingual version, passive recognition (choosing the correct translation) was significantly easier than active recognition (choosing the L2 target word that matched a given L1 word). In the monolingual version, on the other hand, all the recognition items were in L2 only, which may have contributed to a similar difficulty. In spite of this difference between the monolingual and the bilingual versions of CATSS, it has been demonstrated through the difficulty hierarchy of test items that the ability to recognize words, whether passively or actively, will generally precede the ability to recall them, and that recall of meaning will precede the recall of form.

In the monolingual study, the trial sample was made up entirely of learners of English as a second language (ESL) residing in New Zealand with ample opportunities to use English actively in both speech and writing. It is noteworthy that, in our study, an almost identical pattern of performance and progress emerged with EFL learners, whose main exposure to English is in the classroom. This similarity provides additional support for the construct of the strength of vocabulary knowledge.

In addition to validating the strength hierarchy, we wanted to find out which component of knowledge was most strongly associated with classroom success in L2 performance. Each one of the four strength modalities was found to correlate significantly with the learner's English class grade, but the highest correlation (.65) was with passive recall, that is, with the ability to retrieve the meaning for given words. Regression analysis showed that 42.6% of the total variance in the class grade could be explained by the vocabulary knowledge investigated here, that is, the knowledge of form-meaning link. Bearing in

mind that language class grade reflects performance on reading, listening, speaking and writing, grammatical accuracy, sociolinguistic appropriateness, and language fluency, the above figure indicates an impressive contribution of a solid lexical knowledge to overall language success. As stated earlier, vocabulary has been shown to be associated with reading, writing, and academic success. In this study, we isolated the element of word knowledge to which this success is most related: the ability to recall the meaning of words. Recall of meaning is crucial in comprehension. Possibly, the students in our sample, that is, those in a foreign language context, are evaluated primarily on comprehension skills in school. Correlation results may be different in an ESL academic context, in which, in addition to comprehension requirements, students are expected to demonstrate a high level of spoken and written expression, or in an English for specific purposes context, in which learners are specifically instructed to produce the language, as in the case of, for example, English for tourism.

This study did not set out to investigate the proportion of frequent and infrequent vocabulary in L2 learners' lexicon or the growth of L2 vocabulary over time. Still, our test results provided interesting data about possible characteristics of word knowledge development. At the onset of the experiment, we matched specific tests (2,000-word, 3,000-word, 5,000-word frequency levels, AWL) with specific L2 proficiency levels of the participants, as specified in Table 6. The more advanced the students were, the more difficult test they received.⁸ The results of the passive-recognition subtest presented in Table 7 show that the match was indeed successful, since the scores on the 2nd-, 3rd-, and 5th-thousand word levels (27.79, 27.80, 28.60, out of 30, respectively) were nearly perfect and, on a subsequent *F*-test, did not appear significantly different from one another.⁹ Each of the three groups of learners tested had very good passive recognition of 2,000, 3,000, and 5,000 words, respectively. It is the other scores in Table 7 that can shed light on vocabulary development. If vocabulary development

progressed along all four modalities in the same way, the three groups should have received similar scores to one another on the other three subtests (active recall, passive recall, and active recognition). The active-recognition score might have been close to 26.53, passive recall close to 18.6, and active recall close to 6.51. However, the table shows that the scores of the groups decreased through the frequency levels 2nd thousand, 3rd thousand, and 5th thousand. Moreover, a subsequent analysis (Tukey's post hoc test) revealed that, in each of the three modalities, the three learner groups were significantly different from one another.

In terms of vocabulary development, this suggests that an increase in learners' ability to recognize the meaning of words, from 2,000 words to 3,000 or 5,000 words, is accompanied by a smaller increase in recognition of form, an even smaller increase in recall of meaning, and a much smaller increase in recall of form. Put differently, a small increase in the ability to recall words productively (active recall) would imply a much larger increase in passive-recall vocabulary and even larger increase in recognition vocabulary. These results are in accordance with Laufer (1998) and Laufer and Paribakht (1998), who showed that three types of vocabulary knowledge (comprehension, elicited production, and free production) developed in different ways.

Conclusion

In this article we described a bilingual computerized test of vocabulary size and strength and a paper-and-pencil study in which we investigated the validity of the assumptions of the strength hierarchy, that is, that the ability to recognize words, whether passively or actively, generally precedes the ability to recall them, and that retrieval of meaning precedes retrieval of form. The other question addressed in the study was which type of strength of knowledge was best related to general language competence. Results showed that the assumed sequence of the four strength modalities was present at all word frequency

levels. This supports our basic hypothesis that learners' knowledge of form-meaning relationship can be construed as a hierarchy of four degrees of strength: active recall, passive recall, active recognition, and passive recognition. Moreover, results showed that increase in vocabulary size may occur differently at different strength modalities. As for the relationship between the four strength components of knowledge and classroom success in foreign language, we found an impressive contribution of vocabulary, 42.6%, to the total variance in grades in foreign language class. The best predictor of class success was passive recall of meaning.

The study showed that knowledge of form-meaning link is not an all-or-nothing phenomenon but depends on what the learner is required to do with the knowledge: recall the form, recall the meaning, recognize the form, or recognize the meaning. Existing tests of vocabulary size have not been able to capture this distinction. CATSS, a test of vocabulary size and strength, has the capacity to provide this crucial information for researchers and educators.

Revised version accepted 7 January 2004

Notes

¹Active knowledge will be distinguished from "use," which is the willingness of the learner to put the knowledge to use.

²We do not refer to our test as a test of size and *depth*, since the term *depth* may include word features (e.g., grammatical properties, collocations, pronunciation) in addition to meaning.

³First trials with the computerized version of CATSS, with a small sample of learners and with the jump-level and end-of-test functions intuitively determined, showed that the entire test took approximately 30 min to complete.

⁴The native speakers of Russian who participated in this study were attending schools in which Hebrew was the language of instruction.

⁵This is in contrast to the procedure for the computer-adaptive version, in which only the active-recall modality always presents the test taker with 30 items, and the number of items presented in any of the other strength modalities will depend on how many items a candidate has gotten wrong at the previous strength modality.

⁶In order to check whether a different order of presentation had an effect on the results, two groups of 35 participants each were compared prior to the experiment. One group took the passive recall as the second subtest before active recognition, and the other group took the active-recognition subtest before passive recall. We compared the mean scores of the groups on the four subtests for the strengths of knowledge. There were no significant differences between them on any of the subtests. Hence, splitting the participants did not introduce a between-groups factor.

⁷Though our sample consisted of speakers of three different L1s, we present the results for the entire sample. When we separated the results of the three L1 groups, we found that each of the groups displayed the same hierarchy of the four strengths of meaning.

⁸Low-frequency words are considered more difficult than high-frequency words, since low-frequency words are acquired later. Exceptions are words that are cognates in already familiar languages and (in the case of learning a foreign language for specific professional or academic purposes) specialized words.

⁹We will not discuss the Academic Word List, as it includes words of various levels of frequencies (Coxhead, 2000).

References

- Arnaud, P. (1992). Objective lexical and grammatical characteristics of L2 written compositions and the validity of separate component tests. In P. Arnaud & H. Bejoint (Eds.), *Vocabulary and applied linguistics* (pp. 133–145). London: Macmillan.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.
- Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). Cambridge, England: Cambridge University Press.
- Coxhead, A. (2000). A new “Academic Word List.” *TESOL Quarterly*, 34, 213–238.
- Faerch, K., Haastrup, K., & Phillipson, R. (1984). *Learner language and language learning*. Clevedon, England: Multilingual Matters.
- Hatch, E. M., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. New York: Newbury House.
- Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, 21, 303–317.
- Laufer, B. (1997). The lexical plight in second language reading: Words you don’t know, words you think you know and words you can’t guess. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 20–34). Cambridge, England: Cambridge University Press.

- Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics*, 12, 255–271.
- Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, 21, 202–226.
- Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16, 307–322.
- Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, 16, 33–51.
- Laufer, B., & Nation, P. (2001). Passive vocabulary size and speed of meaning recognition: Are they related? In S. Foster-Cohen & A. Nizgorodcew, *EUROSLA yearbook* (pp. 7–28). Amsterdam and Philadelphia: John Benjamins.
- Laufer, B., & Paribakht, T. S. (1998). Relationship between passive and active vocabularies: Effects of language learning context. *Language Learning*, 48, 365–391.
- Meara, P. (1990). A note on passive vocabulary. *Second Language Research*, 6, 150–154.
- Meara, P. (1996, August). *The third dimension of vocabulary knowledge*. Paper presented at the 11th AILA World Congress, Jyväskylä, Finland.
- Meara, P., & Buxton, P. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, 4, 142–151.
- Melka Teichroew, F. J. (1982). Receptive versus productive vocabulary: A survey. *Interlanguage Studies Bulletin*, 6, 5–33.
- Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines*, 5, 12–25.
- Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Rowley, MA: Newbury House.
- Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge, England: Cambridge University Press.
- Palmberg, R. (1987). Patterns of vocabulary development in foreign language learners. *Studies in Second Language Acquisition*, 9, 202–221.
- Read, J. (1989). *Towards a deeper assessment of vocabulary knowledge*. Paper presented at the Eighth World Congress of Applied Linguistics, Sydney, 1987. (ERIC Document Reproduction Service No. ED301048)
- Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, 10, 355–371.
- Read, J. (1997). Vocabulary and testing. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 303–320). Cambridge, England: Cambridge University Press.
- Read, J. (2000). *Assessing vocabulary*. Cambridge, England: Cambridge University Press.

- Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing*, 18, 1–32.
- Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly*, 10, 77–89.
- Ringbom, H. (1987). *The role of the first language in foreign language learning*. Clevedon, England: Multilingual Matters.
- Saville-Troike, M. (1984). What really matters in second language learning for academic achievement? *TESOL Quarterly*, 18, 199–219.
- Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning*, 48, 281–317.
- Schmitt, N. (1999). The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge. *Language Testing*, 16, 189–216.
- Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, 18, 55–88.
- Takala, S. (1984). *Evaluation of students' knowledge of English vocabulary in Finnish comprehensive school* (Rep. No. 350). Jyvaskyla, Finland: Institute of Educational Research.
- Waring, R. (1997). A comparison of the receptive and productive vocabulary sizes of some second language learners. *Immaculata* (Notre Dame Seishin University, Okayama), 1, 53–68.
- Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review*, 53, 13–40.

Appendix A

2,000-Word Level

Test 1: Active Recall

Translate the following words into English. The first letter of each translation is provided.

S	_____	הרחיב	.1
L	_____	נחמד	.2
A	_____	סידר	.3
D	_____	חוב	.4
E	_____	בחר	.5
B	_____	אמיץ	.6
N	_____	רעש	.7
R	_____	הקלה	.8
V	_____	ניצחון	.9
J	_____	צדק	.10
E	_____	אירוע	.11
S	_____	סוד	.12
F	_____	בשר	.13
E	_____	השכלה	.14
S	_____	סולם	.15
C	_____	קשר	.16
A	_____	מתקפה	.17
T	_____	אוצר	.18
W	_____	הון	.19
U	_____	דחף	.20
I	_____	שיפר	.21
O	_____	מקורי	.22
A	_____	עתיק	.23
E	_____	שלם	.24
E	_____	בחן	.25
P	_____	מקובל	.26
R	_____	שורה	.27
M	_____	ייצר	.28
D	_____	הגן	.29
D	_____	פיתוח	.30

Test 2: Passive Recall

Translate the following words into Hebrew. The first letter of some translations is provided.

Victory	_____	ו
Urge	_____	ד
Arranged	_____	ס
Flesh	_____	ב
Elected	_____	ב
Brave	_____	א
Noise	_____	ך
Relief	_____	ה
Stretched	_____	מ
Debt	_____	ח
Lovely	_____	ו
Event	_____	א
Secret	_____	ס
Justice	_____	צ
Education	_____	ה
Scale	_____	ס
Connected	_____	ק
Attack	_____	מ
Treasure	_____	א
Wealth	_____	ה
Row	_____	ש
Improved	_____	ש
Original	_____	מ
Ancient	_____	ע
Entire	_____	ש
Examined	_____	ב
Popular	_____	מ
Manufactured	_____	י
Defended	_____	ה
Developed	_____	פ

Test 3: Active Recognition

Select the correct translation for each of the following words and circle it.

- | | | | | |
|------------|--------------|----------------|----------------|-------------|
| 1. הרחיב | a. Complain | b. Fix | c. Stretch | d. Hire |
| 2. נחמד | a. Lovely | b. Merry | c. Bitter | d. Slight |
| 3. סידר | a. Lean | b. Prefer | c. Arrange | d. Develop |
| 4. חוב | a. Debt | b. Roar | c. Pride | d. Accident |
| 5. בחר | a. Elect | b. Blame | c. Melt | d. Threaten |
| 6. אמיץ | a. Firm | b. Hungry | c. Local | d. Brave |
| 7. רעש | a. Frame | b. Noise | c. Clerk | d. Wine |
| 8. צדק | a. Coffee | b. Justice | c. Disease | d. Stage |
| 9. הקלה | a. Gift | b. Opportunity | c. Relief | d. Tax |
| 10. ניצחון | a. Victory | b. Birth | c. Operation | d. Game |
| 11. אירוע | a. Wage | b. Skirt | c. Respect | d. Event |
| 12. בשר | a. Crop | b. Profit | c. Flesh | d. Theater |
| 13. השכלה | a. Education | b. Parent | c. Journey | d. Trick |
| 14. סוד | a. Secret | b. Choice | c. Temperature | d. Salary |
| 15. סולם | a. Scale | b. Nail | c. Cap | d. Pupil |
| 16. מתקפה | a. Attack | b. Lack | c. Pen | d. Shadow |
| 17. אוצר | a. Charm | b. Treasure | c. Sacrifice | d. Factory |
| 18. הון | a. Cream | b. Pity | c. Wealth | d. Tip |

19. קישר	a. Wander	b. Recognize	c. Limit	d. Connect
20. דחף	a. Urge	b. Deliver	c. Concern	d. Fold
21. שיפר	a. Burst	b. Improve	c. Adopt	d. Climb
22. עתיק	a. Ancient	b. Curious	c. Difficult	d. Holy
23. שלם	a. Social	b. Sorry	c. Entire	d. Clever
24. בחן	a. Examine	b. Pour	c. Satisfy	d. Surround
25. מקובל	a. Popular	b. Electric	c. Independent	d. Angry
26. מקורי	a. Private	b. Royal	c. Slow	d. Original
27. ייצר	a. Cook	b. Manufacture	c. Owe	d. Seize
28. הגן	a. Provide	b. Open	c. Translate	d. Defend
29. פיתח	a. Try	b. Develop	c. Argue	d. Prove
30. שורה	a. Tip	b. Copy	c. Dust	d. Row

Test 4: Passive Recognition

Select the correct translation for each of the following words and circle it

העסיק	.4	הרחיב	.3	תיקן	.2	התלונן	.1	Stretch .1
מריר	.4	קל	.3	נחמד	.2	שמח	.1	Lovely .2
העדיף	.4	נשען	.3	פיתח	.2	סידר	.1	Arrange .3
חוב	.4	שאגה	.3	גאווה	.2	תאונה	.1	Debt .4
איים	.4	המיס	.3	האשים	.2	בחר	.1	Elect .5
אמיץ	.4	מקומי	.3	רעב	.2	קשוח	.1	Brave .6

						Noise .7
רעש	.4	יין	.3	פקיד	.2	מסגרת .1
						Relief .8
מס	.4	הקלה	.3	הזדמנות	.2	מתנה .1
						Victory .9
משחק	.4	פעולה	.3	לידה	.2	ניחצון .1
						Justice .10
שלב	.4	מחלה	.3	צדק	.2	קפה .1
						Event .11
שכר	.4	כבוד	.3	הצאית	.2	אירוע .1
						Secret .12
משכורת	.4	חום	.3	בחירה	.2	סוד .1
						Flesh .13
תיאטרון	.4	בשר	.3	רווח	.2	יבול .1
						Education .14
תכסיס	.4	מסע	.3	הורה	.2	השכלה .1
						Scale .15
תלמיד	.4	כובע	.3	מסמר	.2	סולם .1
						Connect .16
קישר	.4	הגביל	.3	הכיר	.2	שוטט .1
						Attack .17
צל	.4	עט	.3	חוסר	.2	מתקפה .1
						Treasure .18
מפעל	.4	הקרבה	.3	אוצר	.2	קסם .1
						Wealth .19
קצה	.4	הון	.3	רחמים	.2	קצפת .1
						Urge .20
קיפל	.4	הדאיג	.3	העביר	.2	דחף .1
						Improve .21
טיפס	.4	אימץ	.3	שיפר	.2	פרץ .1
						Original .22
מקורי	.4	איטי	.3	מלאכותי	.2	פרטי .1
						Ancient .23
קדוש	.4	קשה	.3	סקרן	.2	עתיק .1
						Entire .24
חכם	.4	שלם	.3	מצטער	.2	חברתי .1
						Examine .25
הקיף	.4	סיפק	.3	שפך	.2	חקר .1
						Popular .26
עצבני	.4	עצמאי	.3	חשמלי	.2	מקובל .1
						Row .27

אבק	.4	העתק	.3	קצה	.2	שורה	.1
						Manufacture	.28
תפס	.4	חייב	.3	ייצר	.2	בישל	.1
						Defend	.29
הגן	.4	תרגם	.3	פתח	.2	סיפק	.1
						Develop	.30
הוכיח	.4	התווכח	.3	פיתח	.2	ניסה	.1

Appendix B

Responses ($n = 30$) for the Word *tradition*

Student number	Active recall	Passive recall	Active recognition	Passive recognition	Passive recognition
1	—	+	+	+	+
2	+	+	+	+	+
3	—	+	+	+	+
4	—	—	+	+	+
5	+	+	+	+	+
6	+	+	+	+	+
7	+	+	+	+	+
8	+	+	+	+	+
9	+	+	+	+	+
10	+	+	+	+	+
11	+	+	+	+	+
12	+	+	+	+	+
13	—	+	+	+	+
14	—	+	+	+	+
15	+	⊖	+	+	+
16	—	+	+	+	+
17	—	+	+	+	+
18	—	+	+	+	+
19	—	+	+	+	+
20	—	—	+	+	+
21	—	+	+	+	+
22	—	+	+	+	+
23	+	+	+	+	+

Note. Circled items represent deviations from the scale.

Appendix B (continued)

Student number	Active recall	Passive recall	Active recognition	Passive recognition	Passive recognition
24	+	+	+	+	+
25	+	+	+	+	+
26	-	+	+	+	+
27	-	+	+	+	+
28	-	+	+	+	+
29	+	+	+	+	+
30	-	+	+	+	+

Copyright of Language Learning is the property of Blackwell Publishing Limited and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.