Language Testing

http://ltj.sagepub.com

The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge

Norbert Schmitt

Language Testing 1999; 16; 189
DOI: 10.1177/026553229901600204

The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge

Norbert Schmitt *University of Nottingham*

In this paper the author argues that issues of construct validity should be given more prominence in the validation of lexical test items. One way of determining the construct validity of vocabulary items is to interview subjects directly after taking the items to ascertain what is actually known about the target words in question. This approach was combined with the framework of lexical competency proposed by Nation (1990) in an exploratory study which investigated the behaviour of lexical items on TOEFL. In individual interviews, six TOEFL vocabulary items were given to 30 pre-university international students who were then questioned about their knowledge of the target words' associations, grammatical properties, collocations and various meaning senses. The results suggest that the type of item currently employed in TOEFL does not adequately reflect association, grammatical and collocational knowledge, and that even meaning knowledge is not captured as well as might be hoped. This indicates that the field could benefit from deeper exploration of what vocabulary test items are actually measuring.

I Introduction

Vocabulary is widely acknowledged as one of the key components necessary for L2 (second-language) proficiency. It is not surprising then that vocabulary has traditionally been one of the language components measured in language tests. This goes back at least to 1916 when Daniel Starch published what Spolsky (1995) believes to be the first modern language tests. More recently, when the *Test of English as a Foreign Language* (TOEFL) was launched in 1964, vocabulary was also included, with its own section (see Read, 1997 for more on the history of vocabulary testing, especially as concerns the TOEFL). Later, a study by Pike (1979) found that the 'words in context' vocabulary items correlated highly with the reading comprehension items also given in the study. The outcome was that vocabulary and reading comprehension items were combined in one section

Address for correspondence: Norbert Schmitt, Department of English Studies, University of Nottingham, Nottingham NG7 2RD, UK; e-mail: Norbert.Schmitt@nottingham.ac.uk

Language Testing 1999 16 (2) 189–216 0265-5322(99)LT168OA © 1999 Arnold

on the TOEFL. As a result of research by Henning (1991), the type of vocabulary item was changed in 1995 to one in which the target words were embedded in the reading comprehension passage with a four-option multiple-choice item attached, such as in the following example:

> The first category of glaciers includes those massive blankets that cover whole continents, appropriately called ice sheets. There must be over 50,000 square kilometers of land covered with ice . . .
>
> The word ‘massive’ in line 4 is closest in meaning to
> (A) huge
> (B) strange
> (C) cold
> (D) recent
>
> (*TOEFL Practice Tests*, 1995: 36)

This type of item has been shown to have good technical characteristics, but the question about what it actually measures still remains. The *TOEFL Test and Score Manual* (1997: 12) explains that Section 3 of the TOEFL, in which the vocabulary items are embedded, attempts to measure mainly reading ability:

> Section 3 measures the ability to read and understand short passages that are similar in topic and style to those that students are likely to encounter in North American colleges and universities. The examinee reads a variety of short passages on academic subjects and answers several questions about each passage. The questions test information that is stated or implied by the passage, as well as knowledge of some of the specific words as they are used in the passage.

However, the last sentence indicates that Section 3 also attempts to measure knowledge of certain words in the passage, presumably through the use of the type of vocabulary item illustrated above. This documentation does specify that the TOEFL focuses on the target words ‘as they are used in the passage’, but does not explain what kind of knowledge is being measured.

In fact, this observation about the deficiency of construct validation can be made not just about these TOEFL vocabulary items, but about most other vocabulary items and tests as well. The implicit assumption seems be that vocabulary items so transparently address the underlying lexical construct that no confirmation of validity is necessary. The present paper questions this view, and argues that we need to have a better understanding of what vocabulary items like those on the TOEFL test actually measure.

II Validity and vocabulary testing

Spolsky (1985) suggests that language testing has evolved through three main phases: a first pre-scientific phase utilizing examinations subjectively marked by a single examiner; a second phase focusing on objectivity and reliability; and a third phase which emphasizes validity along with the aspects in the second phase. Unfortunately, in the case of vocabulary testing, the evolution seems to have largely stopped at the second phase. Studies on vocabulary-item formats, such as Henning (1991) and Schedl *et al.* (1995) have focused on aspects like appropriateness of difficulty, reliability and test speededness. When the issue of validity has been addressed, it has typically not been construct validity (although see Perkins and Linnville, 1987), but other types such as criterion validity, where subparts of a test have been compared to an estimate of total vocabulary size derived from the complete test (Henning, 1991). Construct validity has been largely ignored because most current vocabulary tests are *breadth of knowledge* measures which provide an estimate of *how many* words testees have in their lexicons. This puts the focus on vocabulary size, rather than how well each individual word is known. Items on vocabulary-size tests typically address primarily only the written or spoken form of a word and one of its meaning senses. The appropriacy of such items in developing estimates of vocabulary size has been relatively unquestioned until now.

Of course there is nothing wrong methodologically with extrapolating an estimate of a testee's total vocabulary size from the percentage of the vocabulary sample known on the test. But this process depends on the worth of the individual items from which the total score is extrapolated, highlighting the underlying question of what vocabulary items indicate about knowledge of the word being tested. There is an increasing awareness that measures of vocabulary size alone may no longer be a satisfactory description of vocabulary knowledge, and that there also needs to be a description of *how well* individual words are known (depth of knowledge) (Read, 1993; Schmitt, 1995; Wesche and Paribakht, 1996). The issue of depth of knowledge directly addresses what is known about words. As we shall see, this has clear links with the construct validation of test items.

Following Messick's (1989) lead that construct validity should be considered a unitary concept, Bachman (1990) suggests that it now encompasses the three traditional types of validity: predictive, content and concurrent. Let us look at each of these to explore what they might entail when applied to vocabulary testing.

Predictive validity for some integrated tests can be operationalized relatively easily as the purpose for which the tested language is going to be used. For example, the TOEFL test gives administrators an estimate of the extent to which testees might have language-related difficulties in studying at an English-medium university. However, it is not quite so clear what scores on a vocabulary test could logically predict. While it is true that vocabulary tests can predict success in language-related activities like reading (Laufer, 1992), writing (Laufer and Nation, 1995) and producing correct morphology (Schmitt and Meara, 1997), we must be careful about relating predictive validation for any single language component (like vocabulary) to global language performance, especially considering the numerous other factors which also apply, such as the testee's L1 (first language), motivation and proficiency. In addition, Pike (1979) showed that there is a strong association between scores on the former TOEFL vocabulary and reading comprehension items, which is one reason why vocabulary items have been progressively integrated into the reading section of the test. Thus, in the TOEFL context, the predictive validity of the vocabulary items is bound up with the broader question of what can be inferred from the testee's performance on the reading section as a whole. This makes it considerably more difficult to determine the predictive validity of the vocabulary items themselves.

Most transparently, scores from vocabulary-size items should be used to predict some lexically based language aspect. Unfortunately, not enough is known about L2 vocabulary acquisition to say anything conclusive about the rate or consistency of vocabulary learning. This makes predicting future vocabulary knowledge, such as future vocabulary size, from vocabulary size scores unrealistic at this point in time. In sum, it might be concluded that at the moment predictive validity is not the best way to demonstrate construct validity.

Content validity for individual vocabulary items seems relatively straightforward when compared to other aspects of language. In nondiscrete rule- or regularity-based systems (grammar for instance) it can be difficult to decide which items best represent the system as a whole. On the other hand, vocabulary consists of discrete units, and can thus be explicitly focused on in test items. It does not require expert authority to confirm the content validity of any particular lexical item; the explicit focus should be sufficient in itself.

However, although any individual vocabulary item is likely to have internal content validity, there are broader issues involving the representativeness of the target words chosen. These concern whether those words are representative of not only the lexicon of a language as a whole, but also of language in particular contexts of use. For TOEFL, this means asking questions about the likelihood that the testees will encounter the tested words in their future academic studies, and whether the meaning sense that is the focus of the test item is typical of the way the word is used in an academic context. The first issue can be partially addressed by comparing target words with those on academic word lists such as the university word list (Xue and Nation, 1984). It lists lower-frequency words which were found to be particularly common in a corpus consisting of academic texts from a wide variety of disciplines. Likewise, corpus data can also illustrate which meaning senses are most frequent in academic contexts. A further issue is that vocabulary does not consist only of one-word lexemes, but of multi-word lexemes as well (phrasal verbs, irreversible binomials, idioms, etc.). Since they are part of the lexicon of a language, they presumably need to be considered for inclusion in vocabulary tests. For the TOEFL context, it needs to be determined to what extent such multi-word lexemes occur in academic discourse, both written and spoken. Since all of these broader issues can be addressed relatively satisfactorily by the careful use of corpus procedures which are currently available, it seems that ascertaining content validity need not be a major problem when it comes to vocabulary items.

This leaves concurrent validity, which compares the test scores with those of another measure at roughly the same time. Normally this would entail comparing the sample test's scores with those from an accepted standard test measuring the same construct. The problem in the area of L2 lexis is that there are no accepted standardized tests available for this purpose. The closest things we have to a standardized test are the Levels Test (Nation, 1990) and a number of checklist tests developed by Meara and his associates (Meara and Buxton, 1987; Meara and Jones, 1990; Meara, 1992). Anecdotal evidence seems to indicate that the Levels test works fairly well and it has been used in research studies (e.g. Laufer and Nation, 1995; Schmitt and Meara, 1997). Despite this, it has never been properly validated, which shows that the field definitely needs to focus more attention on validity issues overall. (Note that this validation is now being carried out; see Schmitt and Schmitt (in preparation) and Beglar and Hunt (this issue)). As for the checklist tests, Meara (1996) admits that they are still in a state of development.

The inappropriacy of predictive validation and the lack of concurrent measures seem to limit these approaches as principled methods of establishing the construct validity of vocabulary items. Content validity can be established from corpus evidence, but is probably not sufficient in and of itself. Luckily, there are other options available. In particular, one of the strategies which Messick (1989) proposes for the investigation of validity, qualitative item investigation, seems to have promise here. Chapelle (1994) has already used item analysis as part of an evaluation of C-tests in vocabulary research. However, as always, the question remains how best to operationalize this type of investigation for the targeted vocabulary items, in this case those on the TOEFL. One approach is to address the oft-asked question of what it takes to know a word, and use the answers to construct criteria for a qualitative item investigation. Whatever the answers, acquiring this ‘knowledge’ is an incremental process, so the criteria will have to take into account partial knowledge. This is tantamount to trying to measure *how well* a word is known, i.e. measuring depth of knowledge.

There have been two general approaches to discovering how well words are known: a developmental approach and a dimensions approach (Read, 1997). The developmental approach tries to trace the development of a word’s knowledge over time, usually by means of a scale. The Vocabulary Knowledge Scale (VKS) (Wesche and Paribakht, 1996) is one of the more recent attempts at this approach. Scales like this do get at the incremental issue, but tend to have several problems. Among them are difficulties in describing the number of stages necessary on the scale, defining the stage boundaries, and the fact that there are usually uneven intervals between stages (Read, 1997; Schmitt, 1998a). On the other hand, the dimensions approach attempts to list all of the competencies necessary to use a word in a native-like manner. This approach started with a paper by Richards (1976) and has been followed up by other authors (e.g. Blum-Kulka, 1981; Alexander, 1982). Perhaps the best listing of these competencies has been done by Nation (1990):

1) spoken form of the word
2) written form of the word
3) grammatical behaviour of the word
4) collocational behaviour of the word
5) frequency of the word
6) stylistic register constraints of the word
7) conceptual meaning of the word
8) associations the word has with other related words

These competencies have come to be known as types of *word knowledge*. If it is accepted that these word-knowledge types define what it means to know a word, then one way of determining construct validity is to ascertain how well vocabulary items indicate the mastery of these word-knowledge types for the target words. This word-knowledge framework has already been used in the study of L2 vocabulary acquisition (Schmitt and Meara, 1997; Schmitt, 1998a), and it now seems reasonable to extend it to vocabulary measurement. The general approach is not exclusive to lexis, however; recently Perkins *et al*. (1996) looked at several different domains of general linguistic ability in a similar componential way.

Following this approach, the ideal way to investigate the vocabulary test items qualitatively would be to first give the items, and then to measure what the testees actually know about the target words according to each word-knowledge type. This is clearly unfeasible, both because of the amount of time which would be involved and because adequate measures for some word knowledge types (like register) do not yet exist. But a streamlined version, measuring only selected word-knowledge types, is certainly achievable, and should provide useful information about the construct validity of any test items investigated.

This study is an exploratory attempt to apply this line of reasoning about construct validity to the type of vocabulary items present on the TOEFL test. Subjects are given a number of TOEFL items and are then interviewed to discover what they actually know about the target words' associations, grammatical properties, collocations and various meaning senses. As an exploratory study, hypotheses are not tested, although there is an underlying intuition that just because a test item is answered correctly, it does not mean that everything is known about that word. Similarly, just because an item is missed does not necessarily mean that absolutely nothing is known about the word. These are common-sense intuitions, but it is interesting to see to what extent they hold.

Where the target words are embedded in reading passages in the TOEFL, another issue is raised. Items which test vocabulary in isolation, or with minimal non-defining context, can be seen to be testing previous knowledge of the target words. But where words are embedded in an extended passage, testees will naturally try to guess the word's meaning from the context if it is not previously known. Inferencing from context is a valuable skill, but is a different construct from previous vocabulary knowledge. Should the TOEFL items be seen as measuring this skill or existing knowledge of a word, or both? This study will also attempt to answer this question, at least indirectly.

III Developing the study procedure

Since the main objective of this study is to discover what subjects know about words tested by TOEFL items, the first step was to select a number of items. The reading-comprehension section in which they reside consists of three short passages, each followed by a number of items, some of which are vocabulary items like the example above. Two passages were selected from *TOEFL Practice Tests* (1995) which had three vocabulary items each. The passages and vocabulary items in the *TOEFL Practice Tests* were 'taken from actual test forms given to examinees at two worldwide test administrations' (p. 4), and so can be considered representative of TOEFL items (see Appendix).

Items were chosen which included target words with the greatest number of different meaning senses (*massive*, *peak*, *rare*, *subtle*, *surging* and *trend*).

Previous discussion in this paper has suggested that an important issue of content validity is whether the meaning senses targeted by vocabulary items are typical or not. Since TOEFL relates to an academic context, the meaning senses of its target words should be typical of an academic environment. The researcher referred to the COBUILD Bank of English Corpus (unpublished, but see http://www.cobuild.collins.co.uk/ for more information) and found that the targeted meaning senses for all of the polysemous item words, with the exception of *massive*, were the most frequent ones for both English in general and for English in an academic context. For the item containing *massive*, the meaning sense in the passage seems to be closer to the less-used concrete one of 'physically large, heavy and solid' than the more-frequently-used sense of 'exceptionally large' employed with abstract entities (see Appendix). So we can conclude these TOEFL items have content validity to the extent that they generally focus on typically-used meaning senses.

From previous experience with a similar elicitation design (Schmitt, 1998a) it was clear that eliciting a number of different kinds of word knowledge about target words is time-consuming, so it was decided that only four types would be examined: (1) meaning, (2) word association, (3) collocation and (4) grammatical word class. These particular word-knowledge types were chosen over the other possibilities such as spelling, register or word frequency. This was because previous research had shown that spelling did not seem to pose much of a problem for advanced learners (Schmitt, 1998a), the target words did not seem to have any particularly strong register attributes and intuitions about word frequency seemed less significant than the other word-knowledge dimensions.

The next step was to determine the objective norms and scoring procedures for each of the four word-knowledge types for the six target words. Depending on the word-knowledge type, different sources for these norms were required. In some cases, the norms were already established (meaning and word class), while in others they had to be determined by reference to native norms or corpus data (associations and collocations respectively).

The meaning senses were obtained from three dictionaries: The *Oxford Advanced Learner's Dictionary* (1995), the *Longman Dictionary of English Language and Culture* (1992), and the *COBUILD English Learner's Dictionary* (1989). Multiple meaning senses for each word were accepted which were distinguishable from one another and which the three dictionaries essentially agreed on. Since meaning is the key type of word knowledge, it was measured both productively and receptively. Subjects scored 2 points for productive knowledge of a meaning sense, 1 point for receptive knowledge and no points if the subject was not able to demonstrate sufficient knowledge of a meaning sense once the receptive prompt(s) were given (see below). The maximum score for each word was fixed as the number of meaning senses multiplied by two (2 = productive knowledge). Since the target words had differing numbers of meaning senses (4 maximum, 2 minimum), the meaning results are reported as a proportion. Thus, one meaning sense known productively (2 points), one receptively (1 point) and two unknown (0 points) would equal a meaning proportion for a word of .375 (3 ÷ 8).

Similarly, the dictionaries were consulted for the forms of the four word classes (noun, verb, adjective, adverb). For example, the forms for the target word *massive* were, respectively, *mass* or *massiveness*, [no verb form], *massive* and *massively*. Subjects received credit for each word class for which they could produce a correct form. They were also given credit for cases of a non-existent word class if they indicated that they did not think it existed. Sometimes more than one form for a word class existed (e.g. *rarity* and *rareness* are both noun forms of *rare*); for these, only one form was required for a correct mark. Thus, the scores could range from zero (no word classes known) to 4 (all classes known).

Although there are many association norm lists available (e.g. Postman and Keppel, 1970), most focus on high-frequency words, and so the target words do not appear on these lists. Therefore, the researcher had to gather native associations for these lower-frequency words. Fifty native-speakers gave three association responses each for every target word to form a norming list. The L2 subjects were then given credit for matching associations on this list. Prior association research had given L2 subjects as much credit for matching a unique native response as for the most common response and so, to avoid this, a weighting system was devised. Briefly, it consists of four categories. A score of zero was given to L2 subjects who did not match any of the native norming responses at all. It was found that about 10% of native speakers gave responses that were either unique to themselves or were given by only a few other respondents; thus, they were not representative of the overall native norms. L2 subjects matching only such responses were given a score of 1. Any L2 subject who provided associations which were relatively frequent native responses scored 2 and if the responses were among the most frequent ones they scored 3. Scores of 2 and 3 can be considered native-like. A more detailed explanation of this procedure and empirical support can be found in Schmitt (1998b).

Study in collocations is a relatively new area, and there is no detailed norm listing of collocates for these words. However, computers can be used to calculate word combinations from extremely large corpora, giving us the ability to determine a word's collocates with a great deal of confidence. The collocational norms in this study were arrived at by extracting the most frequent collocates for each target word from the COBUILD Bank of English Corpus (288 million words at the time the research was carried out). It was found that these collocates tended to cluster in semantic fields (Stubbs, 1995) and so the research design took advantage of this fact. The subjects were asked to create sentences which contained the target words in order to see if the sentences also contained any of the collocates obtained from the corpus. The subjects were prompted to create their sentences about topics which addressed the targeted semantic fields. For example, the three sentence topic prompts for *massive* were (1) *If you were talking about war*, (2) *If you were talking about finance or the economy* and (3) *If you were talking about statistics*. Thus, the first prompt was an attempt to elicit collocates on the list like *attack*, *military*, *explosion*, *retaliation* and *strikes*. Getting these prompts right was especially tricky, as they had to be informative enough to suggest possible sentences within a semantic field, but without 'giving away' either the collocates themselves, or the meaning of the target words. The elicited sentences were later scored as collocationally appropriate if one of the collocates from the COBUILD Bank of English norming list occurred in any place within it. Since the subjects were asked for three sentences, the scores could range from zero (no sentence containing a collocate) to 3 (all sentences containing collocates). (See Schmitt, 1998c, for a detailed explanation of this procedure).

IV Subjects

The study involved 30 L2 learners of English. Twenty-seven were international students attending a summer pre-sessional course designed to improve their academic English skills, especially composition writing, in preparation for their entrance into British universities the following October. Three were attending a summer course aimed at improving their general English. The subjects were all students who had either taken the TOEFL test before or who would be the type of student who would take the TOEFL test if they had chosen to study in the United States instead of Britain. They were all volunteers. Their average age was 25.3 (range 18–40), with 16 being male and 14 being female. They came from 8 different countries (9 Japanese, 7 Taiwanese, 7 Thai, 3 Turkish, 1 French, 1 Korean, 1

Omani, 1 Spanish). Fifteen subjects had spent less than one month in English-speaking countries before the study, while the other 14 had spent an average of 4.8 months (range 1–12). One student had spent 4.5 years in English-speaking countries. TOEFL scores were available for 15 subjects (mean 542.6, range 503–610) and IELTS for another 10 (mean 5.4, range 4.5–6.5).

V Interview procedure

The data was collected during interviews held with individual subjects, lasting slightly over an hour on average. The author followed the elicitation instrument (see Appendix) in a lock-step fashion to ensure all interviews would be equivalent. There were no time constraints during the interview, but if an answer was clearly not forthcoming (after perhaps 1–2 minutes with no response), usually in the collocation section, the researcher went on to the next question. After filling in a biodata form, the subject was told that the interview would focus on six words in particular: *massive*, *peak*, *rare*, *subtle*, *surging* and *trend*. They were given a page with these words written in large bold font to refer to and the words were then read to them.

The first task given to them was the association elicitation instrument. The form consisted of the 6 target words, each followed by 3 blanks. The subjects were asked to write in the first 3 words they thought of when they saw the target words. After as many blanks as possible were filled in, the sheet was taken back by the researcher. The second task consisted of taking the TOEFL vocabulary items. This included two passages, both with three vocabulary questions attached. In addition, after each question, the students were asked to check one of two options probing whether the word was previously known, or guessed from context:

______ I know this word
______ I don’t know this word, but guessed from the text

Asking whether the students already knew the word would help shed light on how they went about answering the TOEFL items, particularly in regard to guessing the word’s meaning from the passage context, which relates to what Messick (1989) would call the substantive aspect of construct validity. After finishing, the tests were taken away, and the subjects were not allowed to refer back to them in later tasks.

After confirming that the subjects were comfortable with the metalinguistic concepts *noun*, *verb*, *adjective* and *adverb*, and after giving an example (*stimulate* → *stimulation*, *stimulative*, [no adverb]), the subjects were then questioned about the word class of each target word and asked to verbally give their derivatives if they existed. If there was any doubt about their metalinguistic knowledge, or if subjects seemed unsure during the subsequent task, they were instructed in the grammatical categories. (Research has shown that learners are not necessarily familiar with word classes and their metalinguistic labels, i.e. Alderson *et al*., 1997.) As it happened, most subjects said they knew these concepts very well from their long years of studying English, but to make doubly sure, they were also given lists for each word class which contained numerous sample words which belonged uniquely to that class.

For the fourth task, subjects were asked to verbally compose 3 sentences, in which the target word needed to be included. It was explained that the researcher would be looking for words in these sentences which ‘naturally occurred together’ with the target word, in the same manner that words like *blonde* and *hair* frequently and naturally occur together. They were therefore instructed to give the most common or usual sentence they could think of, rather than an elaborate or creative one. The researcher further explained that since giving *any* sentence with the target words might be difficult to do, he would give them 3 situations or topics to help guide them. The situation-prompt words did not need to be included in the sentence, as they were only to indicate the general realm the sentence should address. If the subject gave a sentence which did not exactly relate to the situation given, but included the target word, it was accepted without further comment.

The last task was to indicate the different meaning senses of the word. All target words were polysemous, a fact made clear to the subjects. The subjects were asked to explain the meaning senses of the words in any manner they could, including giving definitions, using the words in sentences and drawing pictures or graphs to illustrate knowledge of the meaning. (A notepad was available for use throughout the interview.) After the subjects had depleted their productive knowledge of the words, if any meaning senses remained, they were given prompts to help cue any receptive knowledge they may have had. For example, if they knew that *peak* means the top of a mountain, but could not think of anything else productively, they were given the prompts *peak hours* or *peak season* to see if they knew the meaning in this sense. For both productive and receptive meaning knowledge, the subjects’ understanding was probed until the researcher was satisfied that they either knew or did not adequately know a particular meaning sense. After all meanings were exhausted, the interview went on to the next target word, beginning with the association task.

VI Results and discussion

The results from the TOEFL test will be analysed along two main dimensions: (1) whether the appropriate option was chosen and (2) whether the subject indicated that he or she knew the word or whether he or she did not know the word and had guessed its meaning. Thus, there are 4 possible combinations: (1) a known word combined with the correct answer on the test, (2) a known word with an incorrect answer on the test, (3) a word which was guessed with the correct answer given and (4) a guessed word with an incorrect answer given. Table 1 gives the results of the association, word class, collocation and meaning measures for each of these TOEFL result categories.

VII Meaning knowledge and TOEFL responses

Table 1 supports the intuition that if a TOEFL item is correctly answered, it does not indicate that all of the word's meaning senses are completely known. The average meaning proportion score for correctly answered TOEFL items shows that subjects knew about half of the possible meaning content of a word by this study's criteria. This typically meant that the subject knew the most common meaning sense productively and perhaps one or two other meaning senses receptively. If the word had uncommon meaning senses, few of the subjects knew any of these. Using the word *rare* as an example, subjects usually knew the most frequent meaning sense ('unusual' or 'uncommon') productively, and often knew the meaning sense 'lightly cooked meat' either productively or receptively, but seldom knew the senses of 'thin or light air' or 'unusually good, extreme or remarkable' (as in a *rare fright*) at all. So, on average, subjects who

Table 1 TOEFL results vs association, word class, collocation and meaning results ($n = 180$)

TOEFL	Known/ Guessed	*n*	Association (0–3)	Class (0–4)	Collocation (0–3)	Meaning (proportion)
Correct	Known	83	1.386	2.265	1.747	0.540
Correct	Guessed	53	0.566	1.868	0.906	0.319
Correct total		136	1.066	2.110	1.419	0.454
Incorrect	Known	14	1.286	2.286	1.929	0.455
Incorrect	Guessed	30	0.267	1.333	0.767	0.117
Incorrect total		44	0.591	1.636	1.136	0.224
	Known total	97	1.371	2.268	1.773	0.527
	Guessed total	83	0.458	1.675	0.855	0.246

answered a TOEFL item correctly had at minimum a productive knowledge of the most common meaning sense of that word.

For TOEFL items that were incorrectly answered, the subjects had an average meaning proportion score of 0.224. This is nowhere close to 0.000 and indicates that subjects who miss a TOEFL item often have some meaning knowledge of that word. For words with 3 or 4 meaning senses, this would mean that the subject knows about 1 of the meaning senses productively or 2 senses receptively. For words with only 2 meaning senses, it would mean that a subject knows 1 of them receptively. Since if any meaning sense was known it was usually the most common one, this figure suggests that subjects on average knew, at least receptively, the most common meaning sense for the words they had missed on the TOEFL items. An examination of the elicitation instruments shows that this was very often true. For the 44 incorrectly answered TOEFL items, in 19 cases the subjects were able productively to demonstrate knowledge of the most common meaning sense, and in one case receptively. If the results of one particularly weak student were dropped, the number of incorrect TOEFL items drop to 38, which would mean that about half of the remaining subjects (20/38) knew the most common meaning sense of words they missed on the TOEFL.

If we take a meaning proportion score greater than zero to indicate some knowledge of a word's meanings, the figures in Table 2 show that, in this study, when words were at least partially known, the appropriate option was chosen for corresponding TOEFL vocabulary items 83% of the time. While the TOEFL items are reasonably good at distinguishing words which are known, they do not seem to work as well with unknown words. For words which were unknown, as determined by a 0.000 meaning proportion score, about 55% of the corresponding TOEFL items had the appropriate option selected. It was surprising that more than half of the subjects who could demonstrate no knowledge of a word's meanings had been able to answer successfully the relevant TOEFL items a few minutes before. Of

Table 2 Demonstrated meaning knowledge and TOEFL results

TOEFL answers	All words with a meaning proportion greater than zero		All words with a meaning proportion equal to zero	
	n	%	*n*	%
Correct	109	83.2	27	55.1
Incorrect	22	16.8	22	44.9
Total	131	100.0	49	100.0

course the subjects probably selected their answers on the TOEFL items from information inferred from the passages, but in these cases that knowledge must not have been retained, or else it would have shown up on the meaning measure.

This discussion has thus far focused on how well the TOEFL items have indicated overall knowledge of the different meaning senses of a target word. It is also reasonable to ask how well they reveal knowledge of the individual meaning sense indicated by the passage. Table 3 shows that when a TOEFL item was successfully answered, in about 71% of the cases the subject could demonstrate some knowledge of that meaning sense, but in about 29% of the cases could not. When the TOEFL item was unsuccessfully answered, the subjects were still able to demonstrate knowledge of that meaning sense over 45% of the time. In sum, the figures from Tables 1, 2 and 3 suggest that the TOEFL items are less dependable indicators of meaning knowledge than might be hoped, even when it comes to the targeted meaning sense.

VIII Association knowledge and TOEFL responses

From Table 1, we see the average association score is 1.066 for TOEFL items successfully answered and 0.591 for TOEFL items which were missed. The first score indicates a level of association nativeness about the same as one of the native-speaking norming respondents who gave three unique responses. This score suggests that the target word is integrated in the learner's lexicon in a way which is beginning to resemble that of a native speaker who gives non-typical associations, but not yet in a way which resembles the majority of native speakers who give typical associations. The second score indicates that only one or two unique norming associations were matched, suggesting a level of association nativeness which is very minimal. However, the mean figures alone give a slightly misleading

Table 3 Demonstrated knowledge of meaning sense targeted by TOEFL items ($n = 180$)

TOEFL answers	Productive-knowledge demonstrated		Receptive-knowledge demonstrated		No knowledge demonstrated		Total
	n	%	*n*	%	*n*	%	
Correct	85	62.5	12	8.8	39	28.7	136
Incorrect	19	43.2	1	2.3	24	54.5	44

picture, as can be seen from Table 4. In terms of raw numbers, of the 136 correct responses to the TOEFL items, 45 (33%) corresponded with demonstrations of native-like associations, as indicated by scores in categories 2 and 3. But 41% corresponded with demonstrations of no native-like associations at all (category 0). For the incorrect responses to the TOEFL items, 30 of 44 (68%) related to association scores of zero, but 9 of 44 (20%) related to scores in categories 2 and 3.

We can conclude from these results that if a TOEFL vocabulary item is answered correctly, it cannot be taken to mean that the testee associates that word with others in his or her mental lexicon in a native-like way. Perhaps a third of the correctly-answered words might have native-like associations, but an even greater number are unlikely to have native-like associations at all, at least as measured by this type of three-response task. If the TOEFL item is answered incorrectly, we can be more confident that the subject does not have native-like association knowledge for that word, although in about 20% of the cases subjects may indeed possess that knowledge against the indication of the TOEFL item. In short, TOEFL vocabulary items do not seem to be very good indicators of associative word knowledge.

IX Word-class knowledge and TOEFL responses

Table 1 shows that subjects who chose the correct TOEFL option knew about two word-class forms appropriate for the target words. In contrast, subjects who missed the TOEFL item knew the word forms for 1.636 word classes. As these figures are averages, it is again illuminating to examine the raw frequency data, presented in Table 5. The results for correctly-answered TOEFL items indicate a relatively normal distribution, with the mode of 2 and only a few cases corresponding to 0 or 4 word classes. For incorrectly-answered TOEFL items, subjects were unable to indicate a single word class only 16% of the time, and in over half of the cases were able to name two or more word classes. From this, we can see that the subjects

Table 4 Frequency of occurrence in association category ($n = 180$)

	0		1		2		3	
TOEFL answer	*n*	%	*n*	%	*n*	%	*n*	%
Correct	56	41	35	26	25	18	20	15
Incorrect	30	68	5	11	6	14	3	7

Table 5 Frequency of occurrence in word-class category ($n = 180$)

	0		1		2		3		4	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
Correct	4	3	30	22	59	43	33	24	10	7
Incorrect	7	16	14	32	13	29	8	18	2	5

were relatively successful in providing word-class forms. Only in 6% of the cases (11/180) were they unable to give any word class information at all. They were usually able to give the word class of the target word (87%, 157/180), and in cases with only one word class correct, it was usually the word class of the target word as it was presented (70%, 31/44). Usually they were able to give information about one or more of its derivatives as well. In 69% (125/180) of the cases, subjects were able to give the word forms for two or more word classes. Thus, this data indicates that if a subject correctly answered a TOEFL vocabulary item, it was very unlikely that he or she does not know at least the target word's word class, and there was about a 75% chance that he or she knew its form in two or more word classes. If the TOEFL item was missed, there was still only a 16% chance that the word's part-of-speech was not known, with around a 50% chance of two or more word classes being known.

Despite this success, it is interesting to note that only a very small percentage of subjects were able to give all four word classes for the target words, even though they were advanced learners studying to enter British universities. This suggests that learning all members of a word family (similar to developing native-like associations) is something not commonly mastered by L2 learners until relatively late (if ever) in the incremental acquisition process.

X Collocation knowledge and TOEFL responses

The collocation measurement score indicates in how many sentences out of three possible that the subjects were able to include a content word collocate appearing on the list derived from the COBUILD Bank of English corpus. For items which were correctly answered on the TOEFL test, the subjects were able to give collocates for 1.419 sentences on average. For items which were missed, subjects were able to compose 1.136 sentences which included a collocate. The figures in Table 6 indicate that the fact that a TOEFL item was successfully answered does not seem to give any clear indication of collocational ability for that target word. When an item was missed, then

Table 6
Frequency of occurrence in collocational category ($n = 180$)

TOEFL answer	0		1		2		3	
	n	%	*n*	%	*n*	%	*n*	%
Correct	35	26	35	26	40	29	26	19
Incorrect	16	36	9	20	16	36	3	7

it was quite likely (36%) that the subject was unable to create even one sentence including a common collocation, although in 64% of the cases he or she was able to provide one or more sentences. As for associations and word class, we again find that few subjects were able to demonstrate maximum knowledge.

In sum, the figures suggest that TOEFL items do not satisfactorily indicate collocational knowledge. This result may be partially due to the collocation measurement procedure however. Whereas the meaning and word-class measurement procedures were relatively straightforward, and the association procedure built upon a long history of association research, the author believes that the collocation procedure is one of the first attempts at quantifying collocation knowledge. As an experimental procedure, it is unclear to what extent it captures collocational knowledge, but was reported because it is likely to give at least a crude indication of the ability to use a target word's collocates. (See Schmitt, 1998c, for a more detailed discussion of the procedure's limitations.)

XI Knowing a word versus inferencing from context

Of the 83 cases in which the subjects reported they did not know the target word in the TOEFL items and had to guess, 53 of those items were answered correctly. If the subjects were guessing purely at random, we would expect a figure of approximately 21 correct answers from a four-option multiple-choice format. Clearly, the subjects were exceptionally successful in their guesses (64% correct). This success was not due to a number of items being particularly easy to guess. Although the subjects were quite successful guessing the correct response for the item containing *massive* (10 correct guesses, 1 incorrect guess), the other items were not consistently guessed either correctly or incorrectly (*rare*, 3, 4; *subtle* 15, 7; *surging* 16, 11; *trend* 7, 3; *peak* 2, 4). In addition, the success in guessing could not be attributed to a few individuals being exceptional guessers; rather most subjects were relatively successful. All 30 subjects guessed at least one item, but only 7 had more incorrect guesses than correct ones. Moreover, only 4 failed to have at least one successful guess.

The success in guessing could be due to a number of possible factors. It was found that in 40 of the 83 cases, the subjects had nonzero meaning proportion scores. Although the meaning task was done last (and thus may have been affected by some learning during the TOEFL task), this still suggests that the subjects had some prior knowledge of the words, even though they were not confident enough in their knowledge to rate them as known. (The subjects were more accurate at judging which words they did know. In the 97 cases in which the target words were judged as known, 91 had meaning proportion scores greater than zero, indicating at least partial meaning knowledge.) They may also have been unaware of the knowledge they possessed. Another likely reason hinted at above is that they did not know the words beforehand, but inferred their meaning from the attached text. A close look at the passages in question suggests that this might be possible for a person skilled in inferencing. For example, let us look at part of the TOEFL passage paragraph in which *surging* occurs:

> . . . impressive population growth. For every three Canadians in 1945, there were over five in 1966. In September 1966 Canada's population passed the 20 million mark. Most of this *surging* growth came from natural increase. The depression of the 1930s and the war had held back marriages, and the catching-up process began after 1945. The baby boom continued . . .
>
> (*TOEFL Practice Tests*, 1995: 80; my emphasis)

Although the three distractors in the multiple-choice item (*new*, *extra* and *surprising*) are plausible, there are certainly clues available in the above paragraph which could enable testees to come to the appropriate meaning *accelerating*. However, research has shown that the use of such clues is not always a straightforward proposition (e.g. Huckin *et al*., 1993). The contexts in this case seem to be rich enough to enable the inferencing of meaning, but it should be acknowledged that successful guessing from context also requires considerable inferencing skills on the part of the testees.

The design of the study allows a principled, if indirect, examination of the question of guessing from context in the TOEFL passages. In order that the association measurement would not be contaminated by the context given in the TOEFL passages, the association task was given first. Thus it gives an indication of the state of the subjects' knowledge of the word before exposure to the TOEFL test. While the association score of 0.566 for correctly guessed TOEFL items was higher than the score for incorrectly guessed items (0.267), it is still low enough to indicate a minimal knowledge of the words at best. In fact, most of the subjects who correctly guessed the appropriate option did not give even a single native-like association (Table 7), indicating little or no prior knowledge of the word's meaning.

If we allow the assumption that inability to produce native-like associations indicates lack of knowledge of a word, then subjects did not previously know 68% (36/53) of the words for which they were able to chose the appropriate TOEFL option. If the subjects did not have previous knowledge, then the only other source was the TOEFL test itself. Since the multiple-choice items themselves are written in a way not to give away the correct option, then this clearly suggests that the subjects were inferencing the meaning from the passages. In fact, a number of subjects explicitly reported that they did not know the target words, but that they had guessed the meaning from the texts. Although almost everyone agrees that inferencing from context is a positive thing, we may have to reconsider its relationship to the TOEFL vocabulary items. The items may measure existing knowledge about vocabulary in a majority of cases, but this study gives a tentative indication that in a substantial minority (20%, 36/180) of cases the items are measuring inferencing skills.

XII Conclusion

This has been an exploratory study and, as such, some of its limitations need to be acknowledged. Perhaps most importantly, the study included only 6 TOEFL items and 30 subjects, both of which should be considered quite small samples. In addition, the author was exploring new methodologies in the construct validation of vocabulary items. While these seem to hold promise, they cannot yet be considered proven. For these reasons, the results in this paper should be seen as suggestive, rather than conclusive.

As might be expected, the TOEFL vocabulary items were able to give only a limited amount of information about the wider range of word knowledge necessary to master a word. The items were not particularly strong in indicating the subjects' association, word-class and collocation knowledge of the target words. Of course it is not claimed that these items do give this kind of information, although from the rather general description given in the TOEFL documentation, it is

Table 7 Number of items in each association category for guessed TOEFL items ($n = 83$)

	0	1	2	3
Guessed/Correct	36	8	6	3
Guessed/Incorrect	26	1	2	1

difficult to determine precisely what is claimed. A look at the vocabulary items themselves would suggest that each is targeting knowledge of the particular meaning sense used in the corresponding reading passage. But even for the targeted meaning sense, the items were not as robust as one would expect. The upshot is that this exploratory study suggests that we need to have a closer look at what items like this are really measuring. This implies that future item-format validation should include a prominent construct-validity element.

Acknowledgements

The author would like to thank the three anonymous reviewers for their very helpful comments. This research was partially supported by an ORS award (Overseas Research Students Award Scheme) from the Committee of Vice-Chancellors and Principals of the Universities of the United Kingdom (CVCP).

XIII References

Alderson, J.C., Clapham, C.M. and **Steel, D.** 1997: Metalinguistic knowledge, language aptitude, and language proficiency. *Language Teaching Research* 1, 93–121.

Alexander, R. 1982: What's in a four-letter word? Word meaning in English and second language learning. *Die Neueren Sprachen* 81, 219–24.

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

Blum-Kulka, S. 1981: Learning to use words: acquiring semantic competence in a second language. In Nahir, M., editor, *Hebrew Teaching and Applied Linguistics*. Washington DC: University Press of America.

Chapelle, C.A. 1994: Are C-tests valid measures for L2 vocabulary research? *Second Language Research* 10, 157–87.

COBUILD English Learner's Dictionary 1989: London: Collins.

Henning, G. 1991: *A study of the effects of contextualization and familiarization on responses to the TOEFL vocabulary test items*. TOEFL Research Reports, 35. Princeton, NJ: Educational Testing Service.

Huckin, T., Haynes, M. and **Coady, J.** 1993: *Second Language Reading and Vocabulary Learning*. Norwood, NJ: Ablex.

Laufer, B. 1992: Reading in a foreign language: how does L2 lexical knowledge interact with the reader's general academic ability? *Journal of Research in Reading* 15, 95–103.

Laufer, B. and **Nation, P.** 1995: Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics* 16, 307–22.

Longman Dictionary of English Language and Culture 1992: Harlow: Longman.

Meara, P. 1992: *EFL Vocabulary Tests*. Swansea: Centre for Applied Language Studies, University of Wales.

Meara, P. 1996: The dimensions of lexical competence. In Brown, G., Malmkjær, K. and Williams, J., editors, *Competence and performance in language learning*, Cambridge: Cambridge University Press.

Meara, P. and **Buxton, B.** 1987: An alternative to multiple choice vocabulary tests. *Language Testing* 4, 142–51.

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

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

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

Oxford Advanced Learner's Dictionary 1995: Oxford: Oxford University Press.

Perkins, K., Brutten, S.R. and **Gass, S.M.** 1996: An investigation of patterns of discontinuous learning: implications for ESL measurement. *Language Testing* 13, 63–82.

Perkins, K. and **Linnville, S.E.** 1987: A construct definition study of a standardized ESL vocabulary test. *Language Testing* 4, 125–41.

Pike, L.W. 1979: *An evaluation of alternative item formats for testing English as a foreign language*. TOEFL Research Reports, 2. Princeton, NJ: Educational Testing Service.

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

Read, J. 1993: The development of a new measure of L2 vocabulary knowledgeg. *Language Testing* 10, 355–71.

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

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

Schedl, M., Thomas, N. and **Way, W.** 1995: *An investigation of proposed revisions to Section 3 of the TOEFL test*. TOEFL Research Reports, 47. Princeton, NJ: Educational Testing Service.

Schmitt, N. 1995: A fresh approach to vocabulary using a word knowledge framework. *RELC Journal* 26, 86–94.

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

—— 1998b: Quantifying word association responses: what is nativelike? *System* 26, 389–401.

—— 1998c: Measuring collocational knowledge: key issues and an experimental assessment procedure. *ITL Review of Applied Linguistics* 27–47, 119–20.

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

Schmitt, N. and **Schmitt, D.** in preparation: Validating multiple versions of the Vocabulary Levels Test.

Spolsky, B. 1985: What does it mean to know how to use a language? an essay on the theoretical basis of language testing. *Language Testing* 2, 180–91.

—— 1995: *Measured Words*. Oxford: Oxford University Press.

Stubbs, M. 1995: Collocations and semantic profiles: on the cause of the trouble with quantitative studies. *Functions of Language* 2, 1–33.

TOEFL Practice Tests 1995: Princeton, NJ: Educational Testing Service.

TOEFL Test and Score Manual 1997: Princeton, NJ: Educational Testing Service.

Wesche, M. and **Paribakht, T.S.** 1996: Assessing second language vocabulary knowledge: depth versus breadth. *The Canadian Modern Language Review* 53, 13–40.

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

Appendix Prompts and answer key in the study

Association task

Write the first words you think of when you see each prompt word on the line provided.

1. massive ____________ ____________ ____________
2. peak ____________ ____________ ____________
3. rare ____________ ____________ ____________
4. subtle ____________ ____________ ____________
5. surging ____________ ____________ ____________
6. trend ____________ ____________ ____________

Word class, collocation and meaning tasks

MASSIVE

Word class forms

(noun)	*(verb)*	*(adjective)*	*(adverb)*
mass/massiveness	*[none]*	*massive*	*massively*

Collocation sentence prompts

1. If you were talking about war
2. If you were talking about finance or the economy
3. If you were talking about statistics

Meaning senses: prompts (in brackets) and definitions

(building, wall) large + heavy, solid, strong
(crowd, increase) exceptionally large, greater than usual

PEAK

peak	peak	peak/peaked	[none]

1. If you were talking about a business
2. If you were talking about a house
3. If you were talking about geography

(hours, season, sales, output) Point of highest value, intensity, achievement, activity, etc.
(roof, wave, cap) Any shape, edge, or part that becomes narrow and pointed
(geography) the pointed top of a mountain

RARE

rarity/rareness	[none]	rare/rarefied	rarely

1. If you were talking about living things
2. If you were talking about cooking
3. If you were talking about a special person/entertainer

(book, species) unusual, uncommon, one of only a few, not often happening or seen
(steak) lightly cooked meat
(air) thin, light air (as of the mountains)
(fright, gift for comedy, time at a party) unusually good, extreme, or remarkable

SUBTLE

subtlety/subtleness	[none]	subtle	subtly

1. If you are talking about food
2. If you are talking about communication between people
3. If you were talking about a painting (like a Monet)

(flavour, aroma) difficult to detect or describe, fine, delicate
(plan, argument) organized in a clever or complex way, not openly obvious
(mind) able to perceive and describe fine differences, clever in noticing and understanding

SURGING

surge	surge	surging	[none]

1. If you were talking about the natural world
2. If you were talking about business, finance or economics

3. If you were talking about people at a big sports or entertainment event

(crowd, tide) move forward suddenly and powerfully, in a mass or in waves
(sales, anger, electricity) sudden great/powerful increase in something

TREND

trend/trendiness	trend	trendy	trendily

1. If you were talking about economics
2. If you were talking about the clothing industry
3. Any topic, but you must include an adjective which describes the noun *trend* {a(n) ____ trend}

(economic, political, financial) a general tendency or direction
(clothing) a fashion or style

TOEFL VOCABULARY ITEMS

There are two basic types of glaciers, those that flow outward in all directions with little regard for any underlying terrain and those that are confined by terrain to a particular path.

The first category of glaciers includes those massive blankets that cover whole continents, appropriately called ice sheets. There must be over 50,000 square kilometers of land covered with ice for the glacier to qualify as an ice sheet. When portions of an ice sheet spread out over the ocean, they form ice shelves.

About 20,000 years ago the Cordilleran Ice Sheet covered nearly all the mountains in southern Alaska, western Canada, and the western United States. It was about 3 kilometers deep at its thickest point in northern Alberta. Now there are only two sheets left on Earth, those covering Greenland and Antarctica.

Any domelike body of ice that also flows out in all directions but covers less than 50,000 square kilometers is called an ice cap. Although ice caps are rare nowadays, there are a number in northeastern Canada, on Baffin Island, and on the Queen Elizabeth Islands.

The second category of glaciers includes those of a variety of shapes and sizes generally called mountain or alpine glaciers. Mountain glaciers are typically identified by the landform that controls their flow. One form of mountain glacier that resembles an ice cap in that it flows outward in several directions is called an ice field. The difference between an ice field and an ice cap is subtle. Essentially, the flow of an ice field is somewhat controlled by surrounding terrain and thus does not have the domelike shape of a cap. There are several ice fields in the Wrangell, St. Elias, and Chugach mountains of Alaska and northern British Columbia.

Less spectacular than large ice fields are the most common types of mountain glaciers: the cirque and valley glaciers. Cirque glaciers are found in depressions in the surface of the land and have a characteristic circular shape. The ice of valley glaciers, bound by terrain, flows down valleys, curves around their corners, and falls over cliffs.

1. The word ‘massive’ in line 5 is closest in meaning to ______

(A) huge I know this word ______
(B) strange
(C) cold I don’t know this word, ______
(D) recent but guessed from the text

2. The word ‘rare’ in line 19 is closest in meaning to ______

(A) small I know this word ______
(B) unusual
(C) valuable I don’t know this word, ______
(D) widespread but guessed from the text

3. The word ‘subtle’ in line 28 is closest in meaning to ______

(A) slight I know this word ______
(B) common
(C) important I don’t know this word, ______
(D) measurable but guessed from the text

Basic to any understanding of Canada in the 20 years after the Second World War is the country’s impressive population growth. For every three Canadians in 1945, there were over five in 1966. In September 1966 Canada’s population passed the 20 million mark. Most of this surging growth came from natural increase. The depression of the

1930s and the war had held back marriages, and the
catching-up process began after 1945. The baby boom
continued through the decade of the 1950s, producing a
(10) population increase of nearly fifteen percent in the five years
from 1951 to 1956. This rate of increase had been exceeded
only once before in Canada's history, in the decade before
1911, when the prairies were being settled. Undoubtedly, the
good economic conditions of the 1950s supported a growth
(15) in the population, but the expansion also derived from a
trend towards earlier marriages and an increase in the
average size of families. In 1957 the Canadian birth rate
stood at 28 per thousand, one of the highest in the world.

After the peak year of 1957, the birth rate in Canada
(20) began to decline. It continued falling until in 1966 it stood at
the lowest level in 25 years. Partly this decline reflected the
low level of births during the depression and the war, but it
was also caused by changes in Canadian society. Young
people were staying at school longer; more women were
(25) working; young married couples were buying automobiles or
houses before starting families; rising living standards were
cutting down the size of families. It appeared that Canada
was once more falling in step with the trend towards smaller
families that had occurred all through the Western world
(30) since the time of the Industrial Revolution.

Although the growth in Canada's population had slowed
down by 1966 (the increase in the first half of the 1960s was
only nine percent), another large population wave was
coming over the horizon. It would be composed of the
(35) children of the children who were born during the period of
the high birth rate prior to 1957.

4. The word 'surging' in line 5 is closed in meaning to ______

(A) new I know this word ______
(B) extra
(C) accelerating I don't know this word, ______
(D) surprising but guessed from the text

5. The word 'trend' in line 16 is closest in meaning to ______

(A) tendency I know this word ______
(B) aim
(C) growth I don't know this word, ______
(D) directive but guessed from the text

6. The word ‘peak’ in line 19 is closest in meaning to ________

(A) pointed   I know this word ________
(B) dismal
(C) mountain   I don’t know this word, ________
(D) maximum   but guessed from the text