PAPER DETAILS

TITLE: Multiple-Choice Test Items of Foreign Language Vocabulary

AUTHORS: Meral ÖZTÜRK

PAGES: 399-426

ORIGINAL PDF URL: https://dergipark.org.tr/tr/download/article-file/153324

**Eğitim Fakültesi Dergisi**
http://kutuphane. uludag. edu. tr/Univder/uufader. htm

# Multiple-Choice Test Items of Foreign Language Vocabulary

**Meral Öztürk**
*Uludağ Üniversitesi, Eğitim Fakültesi*
*mozturk@uludag.edu.tr*

**Abstract.** This paper reviews the various multiple-choice formats used in testing foreign language vocabulary with special reference to the underlying constructs of vocabulary competence. While all formats are argued to categorically measure recognition of second language word meaning, they are claimed to differ with respect to whether they measure receptive recognition or productive recognition, following a distinction drawn by Nation (2001). A further distinction is made between those formats that measure abstract knowledge of vocabulary and those that measure lexical ability. The paper also discusses problems associated with the contextualization of the target items in the receptive recognition ability formats and considers three proposals for ensuring the processing of the context by the test-taker. It further discusses how receptive formats could be transformed into productive formats by manipulating the relative difficulty of the target and the choice words using word frequency as an index of difficulty.

**Key Words:** Foreign language, vocabulary tests, multiple-choice tests, test formats, contextualization.

**Özet.** Bu makalede yabancı dilde sözcük bilgisinin ölçülmesinde kullanılan çoktan seçmeli soru tipleri incelenmekte ve bu soru tipleriyle sözcük bilgisinin hangi yönlerinin ölçüldüğü irdelenmektedir. Söz konusu soru

tiplerinin tümünde sözcük anlamlarını tanıma becerisinin ölçüldüğü öne sürülerek, bu soru tipleri çeşitli kriterler kullanılarak gruplandırılmaktadır. Makalede ayrıca ölçülmek istenen sözcüklerin bir metin içerisinde sunulmasına ilişkin problemler ve soruların metne dayalı olarak cevaplandırılmasını gerekli kılacak öneriler tartışılmaktadır. Bunun yanısıra, seçeneklerdeki sözcüklerin farklı kullanım sıklığı düzeylerinden seçilmesi suretiyle soru tiplerinin birbirine nasıl dönüştürüleceği açıklanmaktadır.

**Anahtar Kelimeler:** Yabancı dil, sözcük testleri, çoktan seçmeli testler, soru tipleri, metin içerisinde ölçme.

## I. Introduction

Multiple-choice items are popular item types in language testing. They are quick to administer, easy to score, can be applied to a large number of students in a short time, and are highly reliable. Vocabulary multiple-choice tests in addition are so easy to prepare that they can even be automatically produced by computers. Coniam (1997) describes a computer program which generates multiple-choice vocabulary tests that are relatively acceptable within only a couple of minutes.

A number of different formats are available for testing vocabulary through the multiple-choice type. Henning (1991) identifies eight different formats. In this paper, I describe eleven basic formats which can be modified to produce other formats. An issue of crucial importance is describing the kind of word knowledge / ability that is measured by a given format. A correct description of the underlying constructs will have direct relevance to the difficulty of a test as well as to the interpretation of test results in terms of the lexical competence of the learners.

This paper attempts to make a comprehensive and critical review of multiple-choice vocabulary formats with respect to the constructs they measure. First, the constructs that underlie multiple-choice vocabulary items are discussed including those that are common to all formats as well as those that are different between formats. Two dichotomies are offered to distinguish between formats: receptive vs productive and knowledge vs ability. Then, in the following two sections, the various formats are illustrated and evaluated on the basis of these distinctions. Finally, the relative difficulty of the target and choice words is discussed as being decisive in determining the construct underlying a multiple-choice item, and transformation of the measured construct from receptive to productive by changing difficulty is illustrated.

## II. Construct description of multiple-choice vocabulary items

The two constructs that are commonly believed as being tested by a multiple-choice vocabulary item are *word meaning* and *word recognition*. Meaning is used in contrast to other components of word knowledge such as word morphology, word association, collocation, or style, to name but a few, and recognition is used in contrast to recall.

Multiple-choice vocabulary formats in general typically measure only two of the many components of word knowledge: the form and the meaning (see Richards, 1976; Nation, 1990, p. 31; Nation, 2001, p. 27; Ozturk, 2003 for taxonomies of word knowledge components). Even these are tested partially as a written multiple-choice test often measures the knowledge of the written form but not the spoken form; and also of the many meaning senses available for a given word it measures only one (i.e., often the most typical sense).

Schmitt (1999) provided evidence that multiple-choice vocabulary items do not measure lexical knowledge in sufficient depth. He investigated the degree to which a multiple-choice vocabulary item in the TOEFL measured four types of vocabulary knowledge: meaning, word class, collocation and association. Learners first answered multiple-choice vocabulary questions from the TOEFL and then they were tested separately for their knowledge of the same TOEFL words with respect to each knowledge type. The results suggested that "the [TOEFL] items were not strong in indicating the subjects' association, word class and collocation knowledge of the target words" (p. 208-9), although it was a better indicator of meaning knowledge. This suggests that the typical multiple-choice vocabulary item is a measure of word meaning knowledge, which also implies knowledge of the form. This does not mean, however, that the multiple-choice format cannot be extended to test other types of word knowledge as well. The following examples from the British Council and the BBC's web-site 'Teaching English' illustrate how the multiple-choice format can be used to test other components of word knowledge. The item below, for instance, tests collocation knowledge:

***1.*** *Which of the following <u>cannot</u> be delivered?*

*a. a baby*

*b. a letter*

*c. a smile [key]*

*d. a speech*

*(The 'Teaching English' site)*

To answer this item correctly, the learner needs to know that *baby*, *letter* and *speech* each collocate with *deliver*, but *smile* does not. The item below tests the differences between a set of semantically related words:

***2.** The following are all types of money. Which one do you borrow when you want to buy a house?*

*a. pocket money*

*b. mortgage [key]*

*c. allowance*

*d. grant*

*(The 'Teaching English' site)*

The following item requires knowledge of corresponding formal and informal forms:

***3.** Choose the more formal alternative for the word in bold.*

*That picture cost me twenty **quid**.*

*a. dollars*

*b. pounds [key]*

*c. pence*

*d.cents*

*(The 'Teaching English' site)*

Finally, the following item tests connotation knowledge (i.e., cultural associations of words):

***4.** Which bird is the symbol of peace?*

*a. pigeon*

*b. dove [key]*

*c. eagle*

*d. parrot*

*(The 'Teaching English' site)*

These examples indicate that multiple-choice vocabulary item is not limited to testing word meaning only. It has the potential for being used to test word knowledge in greater depth than it is usually credited for. In spite of this, the various formats in common use in vocabulary testing are measures of word meaning. This disuse of the multiple-choice format to test other word knowledge types is understandable given the fact that word meaning is basic to word knowledge in the sense that a word will not be considered known unless its meaning is known. Accordingly, the present review will focus on those formats that measure knowledge of word meaning.

While it is commonly agreed that multiple-choice vocabulary items measure recognition knowledge of L2 vocabulary as opposed to recall, the terms *recognition* and *recall* are not always understood in the same way by different scholars. Read (2000, p. 155-6) defines the terms in the following way: "**Recognition** … means that test-takers are presented with the target word and are asked to show that they understand its meaning, whereas in the case of **recall** they are provided with some stimulus designed to elicit the target word from their memory" (p. 155). An example of a recognition test item is where learners are asked to translate L2 words into their native language while the opposite, translating L1 words into the foreign language, would be a recall item. In this respect, it could be argued that there are no true recall items of multiple-choice vocabulary, as the target word is never produced from memory: it is always there, either in the stem or in the options. On the other hand, Read's distinction between recognition and recall fails to categorize some multiple-choice vocabulary formats. Consider the following example:

***5.*** *He was guilty because he did those things _____.*

*a. both*

*b. noticeably*

*c. intentionally [key]*

*d. absolutely*

*(Henning, 1991, p. 4)*

According to Read's definition, this is not a recall item as the word (i.e., *intentionally*) is given in the options and the learner does not produce it from memory. It is not a recognition item either as the test-taker is not being asked to recognize the meaning for a given word. She/he is being asked to recognize the word itself.

Nation (2001, p. 359) defines these terms somewhat differently: "a recognition vocabulary item format involves the use of choices" whereas "a recall item requires the test-taker to provide the required form or meaning". Nation's and Read's definitions differ with respect to the type of word knowledge required by the item. While Read reserves the term *recall* to the recall of form, Nation's use of the term involves recall of either the form or the meaning. Thus, according to Nation, translation in both ways is of the recall type: in translating to the mother tongue, the test-taker recalls the meaning for an L2 word, which he, then, expresses through a word in his native language whereas translating to the foreign language requires the test-taker to recall the L2 word form for a given meaning expressed in the L1.

In the same way, Read uses the term *recognition* for the recognition of meaning only when the form is given (as in L2 to L1 translation) whereas Nation uses the term for the recognition of either. While Read's definition will label only item 7 below as a recognition item, Nation's definition will classify both item 6, which is identical to item 5 and repeated here for comparison, and item 7 as being of the recognition type.

***6.*** *He was guilty because he did those things _____.*

*a. both*
*b. noticeably*
*c. intentionally [key]*
*d. absolutely*
*(Henning, 1991, p. 4)*

***7.*** *He was guilty because he did those things deliberately.*

*a. both*
*b. noticeably*
*c. intentionally [key]*
*d. absolutely*
*(Henning, 1991, p. 4)*

Item 7 is a recognition item for both Read and Nation because it gives the word in the stem and asks the test-taker to recognize the meaning of this word among other meanings expressed as single words in the choices. On

the other hand, item 6 is not a recognition item for Read because the test-taker does not have to recall the word form as it is given in the choices, but a recognition item for Nation because the test-taker is being asked to recognize the form that corresponds to the meaning illustrated in the stem. Thus, while Read's distinction between recognition and recall fails to apply to some multiple-choice items Nation's distinction between the two categorically classifies all multiple-choice vocabulary items as being of the recognition type.

Following Nation, I define a recognition item as one which involves the recognition of either the form or the meaning of the target word and view recognition, alongside word meaning, as a shared construct in multiple-choice vocabulary test formats reviewed here. Thus, a typical multiple-choice vocabulary item measures recognition knowledge of word meaning. Close scrutiny, however, reveals big differences among these formats in terms of the underlying constructs. The two items above, for instance, are similar in that they are both recognition items in the sense defined above and they both measure the meaning of the word *deliberately*. However, the former of the two seems more difficult suggesting a difference in what is being measured.

Henning (1991) uses the terms *supply* and *matching* to refer to these items respectively. Matching items provide the target word in the stem while supply items contain a blank for the target word in the stem. However, the matching-supply distinction is, as Henning states, based on the task involved in answering an item and thus relates to the format. To capture the difference in underlying constructs I use a distinction drawn by Nation (2001, p. 359) between receptive vs productive items.

Nation (2001, pp. 358-360) defines **receptive** items as those that "involve going from the form of a word to its meaning" (p. 359) and they reflect the way we deal with words in reading or listening. In the receptive use of words, we see or hear a word and recall its meaning. In this process, the word form acts as a stimulus for the meaning. On the other hand, **productive** items "involve going from the meaning to the word form" (p. 359) and involve a similar processing of words as in writing or speaking, where we think of the message (i.e., the meaning) first and then search for the words that will convey this message best. The word's meaning, this time, acts as a stimulus for the form. Nation uses the terms receptive recognition vs productive recognition to refer to the distinction within multiple-choice vocabulary family of formats. The receptive recognition formats give the target word in the stem and require the recognition of the meaning in the choices. Productive recognition formats, on the other hand, give the meaning

in the stem and require the recognition of the word form for this meaning in the choices. Thus, item 7 above presents the test-taker with the word form first in the stem (i.e., *deliberately*) and then with alternative meanings in the choices, and therefore is a receptive item. On the other hand, item 6 presents the meaning first through an illustrative stem sentence and then alternative word forms are offered for this meaning in the choices, and thus it is productive. The receptive-productive distinction is related to the matching-supply distinction mentioned earlier in that matching items are typically, but not universally, receptive and supply items are productive.

Both groups of items have contextualized and decontextualized types. While the decontextualized types measure abstract knowledge of vocabulary, contextualized formats are claimed to measure ability to use words. *Receptive knowledge formats* measure recognition of the meaning for a given word form out of context and *productive knowledge formats* measure recognition of the word form for a given meaning out of context. *Receptive ability formats* are meant to tap comprehension of words in context (e.g., a written text) and *productive ability formats* tap ability to use words in language production. Read (2000) defines comprehension and use as follows: "**Comprehension** … means that learners can understand a word when they encounter it in context while listening or reading, whereas **use** means that the word occurs in their own speech or writing" (p. 156). Although in no multiple-choice format the test-taker is asked to produce language, and no multiple-choice item actually measures use, the productive ability formats require a similar processing of words in language use in that the test-taker needs to consider the grammatical and semantic constraints imposed on the target word by the linguistic context to correctly select the word that will fit in a given context, and thus, they can be said to indirectly measure word use. In what follows, I describe and evaluate the multiple-choice formats in four groups each representing a different construct based on the two distinctions drawn above between receptive vs productive and between knowledge vs ability.

## III. Multiple-Choice Receptive Recognition Formats

In this group of multiple-choice formats, the target word is given in the *stem* either in isolation or embedded within context, and the choices offer alternative meanings for the target. As such these items go from form to meaning and are testing receptive knowledge. The formats in this group differ among themselves with respect to the presence/absence of context surrounding the target word in the stem. The decontextualized formats measure recognition knowledge of word meaning while the contextualized

formats are claimed to measure comprehension ability. These are discussed below.

### 1. Receptive Recognition Knowledge Formats

These provide the target word in isolation in the stem and differ with respect to the way the meaning is expressed in the choices.

#### a. Synonym-matching Format

In this format, the target word appears singly in the stem and the choices are one-word options, one of which is the synonym of the target as exemplified below:

***8.*** *gleam*

*a. gather*

*b. shine [key]*

*c. welcome*

*d. clean*

*(Hughes, 2003, p. 180)*

Although one might argue that this item does not test one but five words simultaneously (*gleam* plus the four words in the choices) and indeed, in other items of the same type this might be the case, the item here tests for only one word. While the target is a low frequency word, all of the options are high frequency. Consequently, the target is difficult but the options are easy 'words that the candidates are expected to know' (Hughes, 2003, p. 181). Thus, the words in the choices function as alternative one-word glosses one of which is the meaning of the target.

#### b. Definition-matching Format

The following format also gives the target word in isolation in the stem, but the choices consist of alternative definitions rather than single words:

***9.*** *loathe means*

*a. dislike intensely [key]*

*b. become seriously ill*

*c. search carefully*

*d. look very angry*

*(Hughes, 2003, p. 181)*

This format has the advantage that the meaning can be defined with more precision and the distractors can be made more challenging. In the above example, 'Dislike intensely' is a more precise description of *loathe* than 'dislike'. Also, each choice was made similarly intense as the key by the use of the adverbs *seriously*, *carefully*, and *very* respectively (Hughes, 2003, p. 181). With such intensification of the distractors, a learner who does not exactly know what *loathe* is but has only a vague idea that it relates to some extreme state might easily be distracted from the correct answer as all distractors refer to an extreme degree of one thing or another.

## 2. Receptive Recognition Ability Formats

In these formats, the target is embedded in context. As such, these formats are assumed to require comprehension of the context for a correct response of the item. As will shortly be seen, however, this assumption is not always tenable and these items often fail to measure test-takers' word comprehension ability which they are claimed to measure and end up testing abstract vocabulary knowledge. Three of the formats described here contextualize the target word in the stem while one format contextualizes the target meaning in the options.

### a. Contextualized-stem Formats

The following three formats contextualize the target in the stem and differ in the amount of context they provide. The *sentence-stem format* below provides the target within a sentence context in the stem rather than in isolation while the choices still contain single words. The sentence illustrates the use of the target in the language:

***10.*** *Nutritionists categorize food into seven basic groups:*

*a. clarify*

*b. grind*

*c. classify [key]*

*d. channel*

*(Hale et al. 1988, p. 67)*

This item is similar to the following item in every respect except that the latter (i.e., *the passage-stem format*) provides a more extended context for the target in the stem:

***11.*** *The first category of glaciers includes those massive blankets that cover whole continents, appropriately called ice sheets. These must be over 50,000 square kilometers of land covered with ice…*

*The word 'massive' in line [1] is closest in meaning to:*

*a. huge [key]*

*b. strange*

*c. cold*

*d.recent*

*(TOEFL Practice Tests, 1995, p. 36 cited in Schmitt, 1999, p. 190)*

Another contextualized-stem format is the *receptive cloze*. In this format, the number of words being tested in a passage is increased by selecting multiple targets. The following example taken from Henning (1991, p. 67) is testing ten target words in a single passage (all underlined and numbered) followed by corresponding multiple-choice items with single word options:

***12.*** *We sometimes take for granted the contributions of science and technology in reducing physical ailments[71] or in providing conveniences like the automobile and the airplane. We tend to forget the technicians working earnestly[72] with dogged[73] determination under conditions that may affect their own health deleteriously[74] to provide us with these advantages. Whether chemists working with frothy[75] chemicals in the isolated adjuncts[76] to their laboratories, or aeronautical engineers wedging[77] strips of some not easily corroded[78] alloy into the frame of a weather satellite, all have contributed to the surge[79] in scientific knowledge. We may never attend a meeting of a scientific society to hear some address[80] on the latest breakthroughs, but we have all benefited from scientific endeavour.*

*71. a. doctors*
*b. diseases [key]*
*c. patients*
*d. livestock*

*72. a. seriously [key]*
*b. cautiously*
*c. secretly*
*d. continually*

*73.a. vague*
*b. inspiring*
*c. fruitless*
*d. persistent [key]*

*74.a. superficially*
*b. dramatically*
*c. adversely [key]*
*d. latently*

*75. a. greasy*
*b. foamy [key]*
*c. fluid*
*d. toxic*

*76. a. additions to [key]*
*b. adaptations of*
*c. advertisements of*
*d. advancements in*

*77. a. cramming*
*b.splicing [key]*
*c. wending*
*d. beating*

*78. a. polished*
*b. softened*
*c. taken over*
*d. worn away [key]*

*79. a. lull*
*b. shift [key]*
*c. jump*
*d. cut*

*80. a. bid*
*b. speech [key]*
*c. envelope*
*d. nomination*

*(Henning, 1991, pp. 67-8)*

This format is the receptive version of the cloze-type which will be described in the next section. In the receptive version, the target words are left in where they are in the passage stem and the choices represent alternative meanings. In the productive version, on the other hand, the targets are replaced by blanks in the passage-stem and appear among the choices.

The idea behind such contextualization of target words is to produce test items that reflect the way words are usually experienced in real life. Words are not normally encountered in isolation in receptive situations but are usually surrounded by context, textual and / or situational. So, it is argued that vocabulary tests should also provide test-takers with context for words so that they should be testing their ability to deal with words in real life situations. Research by Pike (1979) and Henning (1991) gave support to the contextualization of vocabulary. Both of these are TOEFL research and Pike's (1979) study led to the adaptation of a new contextualized vocabulary format where the stem consists of an illustrative sentence in place of the former decontextualized types (i.e., synonym matching and definition completion). Henning (1991) compared nine different multiple-choice vocabulary formats. The decontextualized *synonym-matching* format produced the lowest correlations with overall vocabulary scores and the contextualized *passage-embedded sentence stem/multiple-choice matching format*, such as the one in item 11 above, produced the highest correlations.

The passage stem also proved superior to other contextualized types where the context consisted of a single sentence.

The contextualized formats are claimed to measure lexical ability as opposed to knowledge. There are two problems, however, with the implementation of this claim. One is that the provision of context does not guarantee the processing of the context by the test-taker in all correct responses and does not bring about the intended change in the measured construct. This problem was noted first by Bachman (1986, p. 81 in Read, 2000, p. 142) in relation to the sentence-stem format and by Read (2000, p. 145) and by Banerjee and Clapham (2003, p. 117) in relation to the passage-stem format. A test-taker who already knows what 'massive' is in item 11 above, for instance, may completely bypass the preceding text and still answer the item correctly. A successful answer to the item does not require an understanding of the textual context if the target is previously known to the learner (Read, 1997, pp.306-7). In practice, then, this format might be reduced to synonym matching, which will render all efforts at contextualization useless.

The other problem concerns the way how lexical ability is defined. Receptive lexical ability involves two distinct processing abilities: the ability to understand previously known words in context (i.e., comprehension ability) and the ability to infer the meaning of new words from the information provided in the context (i.e., guessing ability). Research has shown that contextualized items often measure guessing ability rather than comprehension ability. Schmitt (1999), in the study referred to earlier, found that more than half of the learners who reported no previous knowledge of the target words (55%) were able to answer the corresponding TOEFL vocabulary items correctly. This suggests that learners used their ability to infer the meaning from the textual context when their knowledge of the word was lacking. The problem here, however, is that there is no way of knowing if a correct answer is based on previous knowledge or successful guessing (Schmitt, 1999, p.195). Consequently, a contextualized multiple-choice vocabulary test would be misleading, for instance, in calculating learners' vocabulary sizes which is essentially an estimate of known words.

In what follows I discuss three proposals for increasing the demand for the test-taker to process the context even when she/he has previous knowledge of the target. One of these proposals belong to Read (2000, p. 12), and involves the manipulation of the options. Read suggests the options to be designed so that they are all possible meanings of the target. He provides the following example:

***13.*** *Humans have an innate ability to recognize the taste of salt because it provides us with sodium, an element which is essential to life. Although too much salt in our diet may be unhealthy, we must consume a certain amount of it to maintain our wellbeing.*

*What is the meaning of consume in this text?*

*a. use up completely*

*b. eat or drink [key]*

*c. spend wastefully*

*d. destroy*

The word *consume* has all of the meanings given in the choices, and in a decontextualized format, all of the options would be correct. However, in the given context it obtains only one of these meanings (i.e., 'eat or drink' sense). Therefore, Read argues, the test-taker needs to understand the context to identify which one is the correct answer. However, the possibility that the test-taker might give a correct response to this item on the basis of his abstract knowledge of the word *consume* has still not been completely removed with this modification. Abstract knowledge of this word is most likely to include the most typical sense (i.e., 'eat or drink' sense) as research has shown that the typical sense of a word is learnt earlier (Ozturk, 1998; Schmitt, 1998; Verspoor & Lowie, 2003). A learner who knows the most typical meaning of the word but is not aware of the other meanings in the distractors may not be distracted by the 'possible' meanings of the word and might still select the correct option without referring to the text.

A better way to ensure the processing of the context is to manipulate the target sense being tested and test a non-typical meaning of the target word. One way to test a non-typical meaning is to place the word in a context where it is used in a non-typical meaning and ask the test-taker to select the meaning for the word as used in the context. The item will be more challenging if distractors are also 'possible' meanings of the word as suggested by Read (2000). These possible meanings will include both the typical meaning sense and other non-typical meanings. This is exemplified below:

***14.*** *Many opportunities exist for people to work from home and, particularly if you have children to consider, that might be the answer for you. (Livewire, 2005)*

*What does the word consider mean in the text above?*

*a. believe*

*b. care about [key]*

*c. think*

*d. examine*

A correct response in this item will require the processing of the context more strongly than a target used in a typical sense as in Read's *consume* example above. A test-taker who acts on the basis of her / his abstract knowledge of *consider* rather than the context is likely to choose the option (c), i.e., 'think', which is the typical meaning sense of *consider*, and will get it wrong. Therefore, a target used in a non-typical meaning will require an understanding of the context for a correct answer since the typical meaning selected on the basis of abstract knowledge will be incorrect.

Another way to ensure the processing of context is to test word reference. TOEFL's reading comprehension section of the paper-based test does, in fact, include items which require the test-taker to identify the referent of a target word. It is unfortunate, however, that these are not seen as testing vocabulary. Item 15 below taken from a sample TOEFL test is testing the meaning of the word *tradition* as used in the passage rather than an abstract decontextualized meaning of the word. The test-taker has to return to the passage to identify what the word refers to. As such, the test-taker is not being questioned about the word's sense (i.e., what it means), but about its reference (i.e., what it is used for). While the sense of the word *tradition* includes 'customs and practices' in a general manner, the referent of this word in the passage is 'the particular practice of sticking to a common measurement of time'.

***15.*** *Read the following passage:*

*The railroad was not the first institution to impose regularity on society, or to draw attention to the importance of precise timekeeping. For as long as merchants have set out their wares at daybreak and communal festivities have been celebrated, people have been in rough agreement with their neighbors as to the time of day. The value of this tradition is today more apparent than ever. Were it not for public acceptance of a single yardstick of time, social life would be unbearably chaotic: the massive daily transfers of*

*goods, services, and information would proceed in fits and starts; the very fabric of modern society would begin to unravel.*

*In line 6, the phrase ''this tradition'' refers to*

*a. the practice of starting the business day at dawn*

*b. friendly relations between neighbors*

*c. the railroad's reliance on time schedules*

*d. people's agreement on the measurement of time [key]*

Most contextualized items provide glosses of abstract meaning senses in the choices and thus end up testing these abstract senses, which has been the case with the items 10-14. The type of item illustrated by 'this tradition', on the other hand, contextualizes the meaning as well and in a manner close to real life use of words. A reader, for example, will normally use his abstract knowledge of a word to work out what that word is used to mean (i.e., what it refers to) in the text rather than use his understanding of the text to work out the abstract meaning of the word unless the word is unknown and needs to be guessed, which were referred to earlier as comprehension ability and guessing ability respectively. The choices for the *tradition* item above are not abstract meaning senses and a test-taker who already knows the word still has to look at the text to identify for which particular tradition the word is being used for. As such, it measures comprehension ability more successfully than other formats.

**b. Passage-embedded Options Format**

The *passage-embedded options format* (Henning, 1991) is another receptive ability format, and is the reverse of the passage-stem format. In this format, the stem presents the target word in isolation and the options are provided within a textual context (e.g., a sentence) where the options are highlighted by being underlined and lettered as in the following example from Henning (1991, p. 66):

***16.*** *ailments*

*(A) (B) (C) (D)*

*Doctors combat diseases in both human patients and in livestock.*

If written successfully, the semantic relation among the distractors that results from the binding effect of the single-sentence context could make this format rather challenging. It would require precise knowledge of the word, and a learner who only knows that the word belongs to a given topic area (i.e., medicine) will not be able to eliminate the distractors as all distractors will belong to the same topic area as the key. Unsurprisingly, in Henning (1991), this format was one of the most difficult for the learners. It produced the lowest mean scores after the *inference sentence/supply type*. Henning's study also suggested that embedding of the target makes more successful vocabulary items than embedding the options, as the former correlated better with the overall test results, and was more reliable.

Like the passage-stem formats, this format may also fail in contextualization. If the learner already knows the target, he/she may not proceed to read the sentence, in which case the item will be testing abstract knowledge rather than ability. Still worse, the item contextualizes a synonym of the target rather than the target itself and thus measure comprehension of the target only indirectly with the assumption that the context is appropriate for both, which will not always be the case with any two synonyms.

## IV. Multiple-Choice Productive Recognition Formats

In this type of multiple-choice item, the meaning is given in the stem and the corresponding target word is provided in the choices. These items go from the meaning to the form and thus they measure productive recognition of words. This type of item must be more challenging for a test-taker than the receptive type discussed in the previous section as the learner is asked to choose the correct form which is arbitrary and difficult to recall as well as to distinguish from other forms. Henning (1991) provided some support to this claim. Of the eight multiple-choice vocabulary formats he investigated, the supply formats (e.g., short-sentence/supply) turned out to be more difficult than the matching formats (e.g., short-sentence/matching). Although the supply-matching distinction is not identical to the receptive-productive distinction I draw here in that not all productive items are of the supply type (see items 17 and 22 below) and although matching items can be either receptive or productive, matching items are typically receptive and supply items are productive. It follows from this that receptive items would be easier than corresponding productive items.

The choices do not display much variation among the several formats in this group in that they are almost always single words or compounds. The

variation occurs in the way how the meaning is provided in the stem, i.e. with or without surrounding linguistic context.

**1. Productive Recognition Knowledge Formats**

These formats provide a decontextualized verbal or visual definition of the target meaning and thus measure an abstract knowledge of the word's meaning.

**a. Definition-stem Formats**

The two formats in this group require the learner to select, from among given alternatives, the word that best fits the verbal definition provided in the stem. One of the formats provides a traditional dictionary definition of the word in the stem using explicit definition markers like *means* or *meaning*. This format is similar to the definition-matching format discussed earlier. The difference between this and the receptive type (see example 9 above) is that in the latter the learner selects the definition for a given word while in the productive format the learner selects the word for a given definition. This format (i.e., *definition-stem*) is illustrated below:

***17.*** *One word that means to dislike intensely is:*

*a. growl*
*b. screech*
*c. sneer*
*d. loathe [key]*
*(Hughes, 2003, p. 181)*

The other format (i.e., *definition completion*) involves a more user-friendly definition with a blank for the target in the stem sentence:

***18.*** *A_____ is used to eat with.*

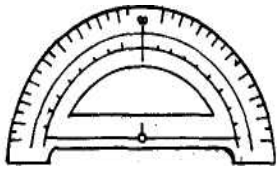
*a. plow*
*b. fork [key]*
*c. hammer*
*d. needle*
*(Read, 1997, p. 305)*

It should be noted, however, that the sentence in the stem is not an illustration of how this word is used in the language. It implicitly defines the word and does not contextualize it.

**b. Picture-stem Format**

Another decontextualized format within the productive type involves completely bypassing language in the stem. Meaning is expressed, in this format, through pictures instead of verbal means as in the following item.

***19.*** *What's this?*



*a. a pencil sharpener*

*b. a set square*

*c. a ruler*

*d. a protractor [key]*

*(Watcyn-Jones, 1994, p. 150)*

In this format, a more direct link is established between the meaning (i.e., concept) and the alternatives. Any disadvantage that may stem from a test-taker's lack of understanding of the verbal description in the stem in other formats who may otherwise know the correct answer will disappear in this format. Among the disadvantages of using picture stems are that they cannot be used with words of a more abstract nature and they will also take up a lot of paper space and will not be very economical to use.

**2. Productive Recognition Ability Formats**

These formats require recognition of the form of the target word corresponding to a meaning illustrated through a verbal context. The two formats in this group differ with respect to the amount of context provided and the number of targets being tested.

**a. Sentence-completion Format**

A contextualized format in the productive category would provide an illustrative sentence in the stem where the slot for the target word is indicated with a blank (i.e., *sentence completion*):

***20.*** *The strong wind _____ the man's efforts to put up the tent.*

*a. disabled*

*b. deranged*

*c. hampered [key]*

*d. regaled*

*(Hughes, 2003, p. 182)*

Unlike the receptive recognition ability type, there is no way for the learner to avoid the context in this format. In order to select the correct word in the options, the learner has to identify the meaning from the sentence first. As such, this item measures ability based on previous knowledge and not guessing ability in that the learner will be able to find the correct answer only if he / she already knows the word disregarding the chance factor. Intelligent guesses based on skilful use of contextual clues are not likely.

**b. Productive Multiple-choice Cloze Format**

*Productive multiple-choice cloze tests* (see item 21 below) will put the target words in larger contexts and therefore are closer to real life use. In this format, words will be deleted from a continuous text on a principled basis (e.g., every fifth, seventh, ninth, etc. word is deleted), and this will be followed by multiple-choice options for each blank.

***21.*** *Australians "Fear Hong Kong Workaholics"*

*Australian bosses disliked hiring Hong Kong employees because they worked too hard and made their Australian colleagues uneasy, ____(1)____ to a recent study in Australia. A University of Wollongong researcher, Ms.Robyn Iredale, commented that a ____(2)____ of the hiring practices of 55 companies also said "there was no ____(3)____ putting a small Asian in a ____(4)____ of authority over taller Australians.". She said: "They said ____(5)____ workers would not like having Asians ____(6)____ because they work too hard….*

*____(1)____*

*a. certainly*

*b. according [key]*

*c. sometimes*

*d. instead*

*e. particularly*

*____(2)____*

*a. driver*

*b. distance*

*c. survey [key]*

*d. dream*

*e. tree*

*____(3)____*
*a. war*
*b. course*
*c. point [key]*
*d. lot*
*e. thing*

*____(4)____*
*a. situation*
*b. letter*
*c. summer*
*d. position [key]*
*e. stage*

*____(5)____*
*a. other [key]*
*b. last*
*c. good*
*d. such*
*e. more*

*____(6)____*
*a. around [key]*
*b. without*
*c. within*
*d. among*
*e. behind*

*(Coniam, 1997, pp. 29-30)*

Cloze tests and multiple-choice cloze tests are often used to measure overall language proficiency, and although vocabulary knowledge is part of this ability and has been shown to contribute considerably to performance on a cloze test (Jonz, 1990), it is difficult to consider a multiple-choice cloze test to be a proper test of vocabulary in this format. The nth word deletion often yields grammatical words (e.g., *around*, *other*, *point*, *according*, in the example above), which do not normally appear in a vocabulary test. On the other hand, a selective deletion of only the content words (Read, 2000, p. 107) or deletion of words that belong to a given grammatical category of content words (e.g. only verbs, only nouns, etc.) or of only those at a given frequency level (e.g., 3,000, 5,000, or 10,000 levels) (Coniam, 1997), is likely to make a more acceptable test of vocabulary. However, the fact that all target words appear in the context of a single text will yield targets, however selected, that are related by topic and representative of only a small part of the L2 vocabulary, which will reduce the content validity of a test designed to measure vocabulary size. On the other hand, this format might be useful as an exercise of vocabulary in a topic area.

## V. Transformation of Receptive Formats into Productive

A number of the receptive formats discussed earlier can be transformed into productive types by manipulating the difficulty of the distractor words relative to the target. The synonym-matching type, for instance, can also be

used to test productive recognition knowledge by reversing the difficulty of the stem word and the option words:

***22.*** *shine*

*a.malm*

*b.gleam [key]*

*c. loam*

*d.snarl*

*(Hughes, 2003, p. 181)*

The difference between this one and the synonym-matching of the receptive type is that, in the productive type, the word in the stem is a higher frequency word than those in the choices while it is the opposite in the receptive type. If we take word frequency as an index of difficulty, which is a common assumption in the field of vocabulary acquisition, the stem word is easier than the choices in the productive type, but more difficult in the receptive type. In the example above, the word in the stem (i.e., *shine*) is not the target. The target is *gleam* and it appears in the choices while *shine* is a synonym of the target. *Gleam* is a more difficult word than *shine* as it is a less common word. Thus, the meaning in this example is given in the stem through a high-frequency word (i.e., *shine*) which could be expected to be known to the test-taker. The distractors are similar to the target (i.e., *gleam*) in being low frequency.

Similarly, contextualized items can also be made productive by changing the relative difficulty of the stem word and choices. Item 11 described as a receptive ability item earlier is transformed below into a productive item by replacing the original target *massive* with a high frequency synonym, i.e. *large*, and the choices with lower frequency words than the target.

***23.*** *The first category of glaciers includes those large blankets that cover whole continents, appropriately called ice sheets. These must be over 50,000 square kilometers of land covered with ice…*

*The word 'large' in line [1] is closest in meaning to:*

*a. massive [key]*

*b. alien*

*c. freezing*

*d. previous*

Unfortunately, many multiple-choice vocabulary items do not observe this difference in difficulty between the stem word and options. The following example from Henning (1991, p. 33) includes both easy and difficult words as options:

***24.*** *The harp is one of the most ancient types of instrument still in use.*

*a. earliest [key]*

*b. strangest*

*c. most expensive*

*d. most intricate*

The options in this example display discrepancies in frequency among themselves and not all are in an appropriate difficulty level. The frequency counts of these words in the BNC corpus is as follows (i.e., higher values represent higher frequency):

***25.*** *ancient (4,910)*

*a. earliest + early (35,384)*

*b. strangest (6,276)*

*c. most expensive (5,746)*

*d. most intricate (518)*

If this item is assumed to be a receptive item where the target is the word *ancient* in the stem, the choices would be expected to be of higher frequency than the target. This is not the case, though, for all of the options. The first option, which is the correct option, is appropriately more frequent than the target. Options (b) and (c) do not seem to be very different from the target in numerical terms. However, when considered in terms of the frequency bands often used in descriptions of vocabulary size (e.g., Nation's Levels Test, 1990), those distractors and the target will fall into different bands, the distractor words being at a higher frequency band than the target. Thus, these two options might be considered acceptable. The last distractor 'intricate', however, is even a less frequent word than the target. Wesche and Paribakht (1996, p.17 in Read (2000, p.78)) argue that such items will, in effect, be testing knowledge of the distractors rather than or in addition to the target. In

some cases, the balance might be so distorted that it might be difficult to say what type of vocabulary knowledge is being tested (i.e., receptive or productive) as in the following example taken from the receptive cloze discussed earlier on p. 410 (the parentheses give the frequency in BNC):

***26.*** *wedging (21)*

*a. cramming (66)*

*b. splicing [key] (88)*

*c. wending (20)*

*d. beating (1817)*

*(Henning, 1991, p. 38)*

Both the target and the options in this example (except *beating*) are low frequency and thus, equally difficult. Does this item, then, test for the meaning of *wedging* (receptive) or does it illustrate the meaning in the passage-stem through the word *wedging* and test for the word *splicing* that has this meaning (productive)? To the extent that we find it difficult to give this answer the item can be argued not to be clear in targeting a particular type of vocabulary construct and thus will be poor in construct validity. On the other hand, a computer program such as the one described in Coniam (1997) can be programmed to select distractors from a higher or a lower frequency band than the target thus creating receptive or productive items as necessary. Also, equating the option words with one another in frequency will enhance the effectiveness of distractors.

It is argued here that the relative difficulty of words in the stem and options should be controlled for in multiple-choice vocabulary items according to whether we are interested in testing receptive vocabulary recognition or productive vocabulary recognition knowledge / ability of the learners. Otherwise, we will fail to test either.

## VI. Conclusion

This paper has discussed the constructs underlying multiple-choice vocabulary formats. All well-known multiple-choice vocabulary formats were described as measuring recognition categorically and word meaning most typically. On the other hand, the various formats that are widely used to test word meaning are argued to differ with respect to the constructs they

measure. Two groups of formats were distinguished: those that measure receptive recognition and those that measure productive recognition. Within each of these groups of formats, those that measure knowledge and ability are distinguished. The problems in contextualization in receptive ability formats have been discussed as well as the possible solutions. Finally, the transformation of receptive types into productive types by changing the relative difficulty of the target and the choices is exemplified.

The improvement of available test formats, including the multiple choice, and development of new ones depends, on the one hand, on an understanding of the limitations of the existing formats in testing communicative ability and sound theoretical descriptions of this ability, on the other. The present paper has been intended as a contribution to the former with respect to multiple choice vocabulary formats. However, descriptions of "communicative lexical ability" (i.e., ability to use words in written and spoken communication receptively and productively) in various situations (both receptive and productive) are still lacking. Chapelle (1988) and Alderson and Banerjee (2002) argue that most work in vocabulary acquisition and testing view lexical competence as knowledge and ignore the differences in the type of vocabulary knowledge required in different contexts of use. A reading activity, for instance, requires knowledge of different aspects of words from a listening activity (e.g., written form and spoken form respectively); an L2 reader will draw on different content vocabulary when reading a scientific article and when reading a novel; different degrees of word knowledge will be required when speaking and reading, etc..

The interactionalist view of construct definition which is currently popular in language testing (Chapelle, 1988) will not only require a description of knowledge traits in specified contexts of use but also a description of the 'strategic competence' (Bachman & Palmer, 1996) of the learner in using his current linguistic knowledge to meet the demands of the context. The strategic vocabulary competence in reading, for example, might involve varying the number of words ignored, guessed from context, or looked up in a dictionary depending on the amount of text words unknown to the learner, the learner's familiarity with the topic and text type, the clarity of the writer's discourse, or the need for detailed understanding of the text. Thus, an L2 learner reading fiction for pleasure can ignore most words that are new whereas an ESP learner reading a textbook chapter in preparation for an exam might look up most, if not all, of the unknown words.

In consequence, vocabulary tests based on an interactionalist construct definition of lexical competence specifying the relevant traits, contexts and strategies will be better equipped to test communicative lexical ability.

## References

Alderson, J.C. and Banerjee, J. 2002. Language testing and assessment (Part 2). *Language Teaching* 35, 79-113.

Banerjee, J., and Clapham, C. 2003. The TOEFL CBT (Computer-based test). *Language Testing* 20 (1), 111-123.

Bachman, L.F. 1986. The Test of English as a Foreigh Language as a measure of communicative competence. In Stansfield, C. W. (Editor) Toward communicative competence testing: Proceedings of the second TOEFL invitational conference. *TOEFL Research Reports*, 21, 69-88.

Bachman, L.F. & A.S. Palmer. 1996. *Language Testing in Practice*. Oxford: Oxford University Press.

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

Coniam, D. 1997. A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests. *CALICO Journal* 14 (2-4), 15-33.

Hale, G.A., Stansfield, C.W., Rock, D.C., Hicks, M.M., Butler, F.A., and Oller, J.W.Jr. 1988. Multiple-choice cloze items and the test of English as a foreign language. *TOEFL Research Reports*, 26. Princeton, NJ: Educational Testing Service.

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

Hughes, A. 2003. *Testing for Language Teachers*. Second Edition. Cambridge: Cambridge University Press.

Jonz, J. 1990. Another turn in the conversation: what does cloze measure? *TESOL Quarterly* 24, 61-83.

Livewire, S. 2005. Why self-employment? Retrieved April 8, 2005 from: http://www.prospects.ac.uk/cms/ShowPage/Home_page/Self_employment/Getting_started/p!emXjF

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

Nation, I.S.P. 1990. *Teaching and Learning Vocabulary*. Boston: Heinle & Heinle.

Ozturk, M. 2003. Lexical competence in the Common European Framework of Reference for Languages. Paper presented at the I. International Symposium on the Common European Framework and Foreign Language Education in Turkey. Bursa.

Ozturk, M. 1998. *Knowing What a Word Means: Acquisition of Noun Polysemy in English by Turkish Learners*. Unpublished doctoral dissertation, University of Reading.

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.

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

Read, J. 1997. Vocabulary and testing. In Schmitt, N. and M. Mc-Carthy (Editors) *Vocabulary: Description, Acquisition, and Pedagogy*. Cambridge: Cambridge University Press, 303-320.

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

Schmitt, N. 1999. The relationship between TOEFL vocabulary items and meaning, association, collocation and word class knowledge. *Language Testing* 16 (2), 189-216.

Schmitt, N. 1998. Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning* 48 (2), 281-317.

*The Teaching English Site*, The British Council & the BBC. (URL:http://www.teachingenglish.org.uk/download/quizzes.shtml#vocabulary)

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

Verspoor, M., & Lowie, W. 2003. Making sense of polysemous words. *Language Learning* 53 (3), 547-586.

Watcyn-Jones, P. 1994. *Target Vocabulary 2*. London: Penguin Books.

Wesche, M., & Paribakht, T.S. 1996. Assessing second language vocabulary knowledge: depth vs. bredth. *Canadian Modern Language Review* 53, 13-39.

## Summary

This paper reviews the various multiple-choice formats used in testing foreign language vocabulary with special reference to the underlying constructs of vocabulary competence. All formats under investigation are argued to categorically measure recognition as opposed to recall. A recognition item is defined here, following Nation (2001), as one which involves the recognition of either the form or the meaning of the target word among alternatives provided when either one is given in the stem. Recall is the retrieval of the meaning or form of the word from memory and is not

involved in the multiple-choice type. Another shared construct among the formats investigated here is word meaning although the multiple-choice format can be extended to measure other word knowledge aspects like word morphology, association, collocation, style, etc.

While the constructs of word meaning and recognition are common to all formats, they are shown here to differ with respect to whether they measure receptive recognition or productive recognition, following a distinction drawn by Nation (2001). The receptive recognition formats give the target word in the stem and require the recognition of the meaning in the choices. Thus, they go from the form of a word to its meaning. Productive recognition formats, on the other hand, give the meaning in the stem and require the recognition of the word form for this meaning in the choices. In contrast to receptive formats, they go from the meaning to the word form.

A further distinction is made between those formats that measure abstract knowledge of vocabulary and those that measure lexical ability. The former measure words in isolation while the latter measure them in textual context.

The paper also discusses problems associated with the contextualization of the target items in the receptive recognition ability formats and considers three proposals for ensuring the processing of the context by the test-taker. One problem is that the provision of context does not guarantee the processing of the context by the test-taker in all correct responses as the test-taker may provide a correct answer for a familiar word without reading the text. As a result this format does not distinguish between answers based on previous knowledge and those based on successful guessing of unknown words. The suggestions to ensure the processing of the context involve using options which are all possible meanings of the target (Read, 2000); testing a non-typical meaning of the target in addition to providing possible meanings in the options; and finally testing word reference (i.e. what the word is used for) rather than abstract meaning senses.

The paper further discusses how receptive formats could be transformed into productive formats by manipulating the relative difficulty of the target and the choice words using word frequency as an index of difficulty. In receptive formats, the options for a vocabulary item are chosen from a higher frequency level than the target and act as glosses for its meaning requiring the test-taker to recognize the meaning for a given form. If the relative frequency of the target and option words is reversed, the item becomes productive as the test-taker will need to go from the meaning provided through a familiar frequent item in the stem to the less frequent word forms in the choices.