English Language Teaching; Vol. 7, No. 10; 2014
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

Two Formats of Word Association Tasks: A Study of Depth of Word Knowledge

Seddighe Jalili Agdam[1] & Karim Sadeghi[1]

[1] Urmia University, Iran

Correspondence: Seddighe Jalili Agdam Karim Sadeghi, Urmia University, Iran. E-mail: Sa194916@gmail.com

Received: April 24, 2014 Accepted: August 26, 2014 Online Published: September 22, 2014

doi:10.5539/elt.v7n10p1 URL: http://dx.doi.org/10.5539/elt.v7n10p1

Abstract

Vocabulary development is an essential goal in any language teaching program, and considering the multidimensional nature of this construct, achieving this goal needs effective assessment of all dimensions of word knowledge, i.e. breadth, depth and accessibility of word knowledge. Most of the current vocabulary assessment tools measure the breadth dimension of vocabulary (Christ, 2011). However, there have been studies which have developed *Selective* or *Productive* word association tasks (WAT) to measure depth of word knowledge (Meara & Fitzpatrick, 2000; Read, 1998). This study used both selective and productive WAT tasks to measure depth of word knowledge in 82 elementary and 71 advanced EFL learners to explore which format is better for assessing deep word knowledge for each group. Results showed that elementary learners did better in selective format while advanced learners acted better in productive format.

Keywords: multidimensionality of word knowledge, depth of word knowledge, word association task (WAT), productive WAT, and selective WAT

1. Introduction

Vocabulary is an essential part of a language. Without a sufficient level of vocabulary knowledge, communication would suffer. When we intend to convey a message, we may be able to do this without the correct form of structure but it would be impossible to convey a message when we do not have at least a basic knowledge of words. Nowadays with the growing use of communicative approach to teaching, the need for a sufficient word list to be accessible when students try to communicate becomes obvious. So EFL learners summarize learning a language in learning its vocabulary by trying to memorize new words. Knowing the importance of this construct (vocabulary knowledge), the need for accurate ways of assessing it becomes clearer to EFL teachers to ensure the efficacy of their teaching method. Unfortunately current methods of assessing this construct are primarily chosen considering economical concerns and convenience. Most of the time vocabulary knowledge is measured through multiple choice formats which requires the test-taker to choose a synonym for a stimulus word or assesses the word knowledge embedded in a reading comprehension test. These ways of testing vocabulary deemphasize the importance of the actual nature of word knowledge and its various dimensions. The aim of this study is to draw the attention of readers and researchers to the vital need for including the exact nature and dimensions of word knowledge in their list of considerations when they want to develop a test to measure vocabulary knowledge. There have been some efforts by vocabulary researchers to define the nature of word knowledge and its different dimensions.

Richards (1976) identifies some aspects of word knowledge as syntactic behavior, associations, constraints, semantic value, usage, different contextual meanings, its morphology and underlying form and derivations. Nation (1990), proposed eight types of word knowledge: 1) the spoken form of a word 2) the written form of a word 3) the grammatical behavior of a word 4) the collocation behavior of the word 5) the frequency of the word 6) the stylistic register constraints of the word, 7) the conceptual meaning of the word and 8) the associations the word has with other words. Beck and McKeown (1991), suggest that vocabulary knowledge should be investigated in terms of levels that range along a continuum from no knowledge to complete knowledge. Chapelle (1998) divided the trait theory of vocabulary knowledge to four dimensions: 1) vocabulary size, 2) knowledge of word characteristics, 3) lexical organization, and 4) process of lexical access. Henriksen (1999) argued that vocabulary knowledge must be approached through 3 dimensions: 1) partial-precise knowledge, 2) depth of knowledge and 3) receptive-productive knowledge of words. Qian (2002), in an attempt to develop a new model to define this construct proposed that vocabulary knowledge consists of four intrinsically connected dimensions: 1) vocabulary size, which refers to the number of words of which a learner has at least some superficial knowledge of meaning; 2) depth of vocabulary knowledge, which includes all lexical characteristics such as phonemic, graphemic, syntactic, semantic, collocational, and phraseological properties, as well as frequency and register; 3) lexical organization, which refers to the storage, connection and representation of words in the mental lexicon of a learner; and 4) automaticity of receptive-productive knowledge, which refers to all the fundamental processes to access the word knowledge for both receptive and productive purposes, including phonological and orthographic encoding and decoding and access to structural and semantic features from the mental lexicon. Laufer, Elder, Hill, and Congdon (2004) have viewed vocabulary knowledge in terms of levels that progress from superficial familiarity to the ability to use new words correctly in speech and writing.

A review of the above frameworks to define the nature of the construct of vocabulary knowledge indicates that, a point of consensus is apparent that they all reject the assessment of word knowledge as a dichotomous wrong or right response. They all agree on multidimensionality of word knowledge. Most conceptions of this dimensionality include two common indices called vocabulary size and depth of vocabulary knowledge. For example researchers such as Read (2000), Qian (2002), and Vermeer (2001) consider vocabulary knowledge as consisting of two dimensions of breadth and depth of vocabulary knowledge. Breadth of vocabulary knowledge has been defined by Nassaji (2004), Qian (2002), and Zareva (2005) as a person's vocabulary size, or approximately the number of words one knows. Without vocabulary breadth understanding sentences or texts is not possible. Some researchers such as Schmitt and Meara (1997), and Wesche and Paribakht (1996), believe that measuring word breadth is of limited value because it ignores the fact that words can be known to a greater or lesser extent. They suggest that any vocabulary test should include a section measuring the depth of word knowledge.

As mentioned above, depth of word knowledge is considered as the second dimension of vocabulary knowledge. This dimension has been an overlooked area of research. Despite being a neglected area of study, it is nonetheless important especially for bilingual learners who are dealing with two languages at the same time. Lexical depth involves many aspects related to the development of literacy skills. Schmitt (2000) defined depth of word knowledge as syntactic properties, possible collocations, pragmatic rules and semantic representation of the words or concepts. Snow and Locke (2001) divided lexical depth to: morphological structure, phonological representation and orthographic representation. Wesche and Paribakht (1996) defined deep word knowledge as the richness of the representation of the known words or how well one knows about words or concepts. Snow (2001), on the other hand suggested that lexical depth includes a receptive and a productive aspect in the sense that learners may have a receptive knowledge of the meaning of a word and recognize its meaning in a text or discourse and be able to choose its collocations and paradigmatically/syntagmatically related words from among a set of given words but not be able to produce such related words when they are confronted with the stimuli word. There have been some efforts to develop tools to assess these two aspects of lexical depth. Most of these tools are based on word association theories and are called Word Asoociation Tasks (WAT).Selective WAT tasks were developed to assess receptive aspect and productive WAT tasks were developed to assess the productive aspect of deep word knowledge.

Read (2004) identified three paths in operationalizing the concept of deep word knowledge:

1) The difference between having a limited unclear idea of what a word means and having much more specific knowledge of its meaning, which is called *precision of meaning*.

2) Knowing the semantic feature of a word and its orthographic, phonological, morphological, syntactic, collocational and pragmatic characteristics which is called: *comprehensive word knowledge*.

3) The incorporation of the word into its related words in the schemata, and the ability to distinguish its meaning and use from related words, which is called *network knowledge*.

This study uses the third line of definition to assess depth of word knowledge in elementary and advanced EFL learners. This line of thought to define depth of word knowledge explores the development of links between sets of words in the mental lexicon. The assumption is that, when a learner increases his vocabulary size, the new words to be learned need to be related and attached to a network of already known words, and some restructuring of the network may be needed as a result. This means that depth of word knowledge can be understood in terms of learner's developing ability to distinguish semantically related words and more generally, their knowledge of the various ways in which individual words are linked to each other that is 'word associations'. By defining depth of word knowledge as the extent and quality of word associations in mind, word association theories become the basis for developing many tests to measure depth of word knowledge called Word Association Tasks

(WAT). As Schmitt (1998) states the elicitation of word associations is a relatively simple procedure, which is one of its attractions. Traditionally, subjects were given a stimulus word and asked to produce the first response which came to mind. For him, the use of word associations holds a great deal of promise in the areas of L2 vocabulary research and measurement. He further claims that word association procedures can be used as an alternative way to test vocabulary.

2. Literature Review

Language knowledge is not acquired in an all-or-nothing fashion (Pearson, Hiebert, & Kamil, 2007). It is incrementally learnt (Nagy & Scott, 2000) so we can not measure knowledge of a word as simply wrong or right and current tools are inadequate for assessing children's vocabulary knowledge (Christ, 2011). Consider the following example taken from Christ (2011, p. 131)

Teacher asks: Jose Do you know what is the meaning of word *Appear*?

Jose says: yes. *Appear* means that like there would be a wizard, and it would be here because magic potion would go onto it.

Teacher: have you ever *appeared*?

Jose: No I don't have a scientist in my house. Just scientists can make things appear.

Here we cannot easily judge whether Jose knows what the word *appear* means or not. He knows some aspect of its meaning, but he does not have enough deep knowledge of it to grasp multiple meanings of this word. We can conclude that the child has established a stage of word knowledge which needs support from teacher to deepen it. Early vocabulary learning needs a purposeful support and to gain this goal, assessing the quality of knowledge children gain about meanings of words is necessary (Verhallen & Schoonen, 1998). Only by utilizing appropriate testing tools can researchers and teachers evaluate the effectiveness of the methods used to teach vocabulary. That is why some vocabulary researchers try to use word association tasks (WAT) to measure not only superficial knowledge of a word but also the depth of that knowledge. Because when a language program ignores the depth of word knowledge gained by children "the arrears of linguistically less proficient children may go unnoticed, and thus pose undesirable educational risks" (Verhallen & Schoonen, 1998, p. 467). Without an appropriate tool for assessing deep word knowledge teachers may only be concerned about whether words are known or not rather than considering the quality or depth of learners' word knowledge. Even to begin any new language course teachers need to assess kinds of knowledge learners have about word meanings because they should build the new material and the new instructions on what learners already know and this is not possible with relying only on multiple-choice vocabulary tests or any other tool which give a superficial information about test-takers quality of word knowledge. Cloze tests were the alternative to check whether a test-taker has the ability to identify the different uses and different functions of a word by determining which word is appropriate for the special context provided by the cloze text. But this tool is not appropriate for young EFL learners who have a basic knowledge of words and a different perception of word meanings. So there is a crucial need for exploring an appropriate tool for assessing vocabulary knowledge.

Read (1993), developed a selective WAT task to measure depth of word knowledge in university students. His initial explorations about depth of word knowledge began with an interview procedure (Read, 1989) which asked test-takers to express orally their knowledge about different lexical aspects of a word. He found that this method is time-consuming and just a few words can be included as stimulus words in such an interview. He was seeking a way to develop a test which has the potential of testing a broad coverage of words. He knew that previous researchers had experimented the free word association tasks for this purpose and had found that it was not a good measure of second language proficiency because the responses were so diverse to such tests and because it had problems with objectivity and scoring reliability. So he developed his selective format (WAT) based on a suggestion by Paul Meara (1983), who thought of presenting test-takers with a stimulus word followed by a set of other words some of which are related to the stimulus and some are not. His test consisted of 40 items each containing a stimulus word followed by 8 other words, 4 of which were related to the target word and 4 were not. Just 4 words were allowed to be selected from among the 8 words. A sample item of his test is as follows:

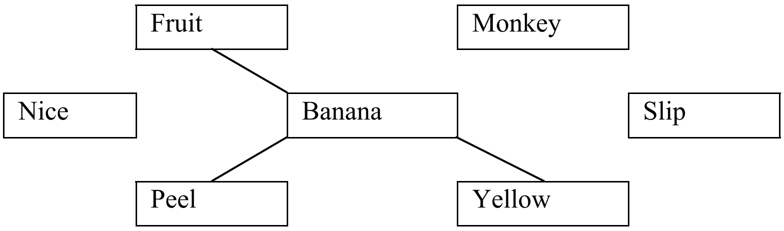
Edit

Arithmetic film pole publishing

revise risk surface text

The stimulus words were selected from University Word List (Nation, 1990) and the distractors were checked to be in the same level of frequency as the stimulus word and were never more difficult than it. Before administering the test to the target group, it was piloted by giving it to ten native-speakers to ensure the researchers' judgment about the associate words and distracters. Then it was administered to two groups of students who were studying in the summer English Proficiency course at the English Language Institute of Victoria University of Wellington in London and had studied Advanced English Vocabulary series by Bernard (1971) which covers a large percentage of the university vocabulary words lists. He used both IRT model and parallel-test method to calculate the reliability of the test and arrived at a good degree of reliability (R = 0.97, P > 0/01) and found it to be a good measure of depth of word knowledge.

After this work by Read (1993) Verhallen and Schoonen (1998) who had used definition task and structured interview to elicit word associations decided to develop a selective task similar to WAT but with level adaptation for elementary learners. To make the test easier, they provided 6 options instead of 8 options, and distractors were also semantically related to true answers. To make it eye-catching for children, the stimulus word was presented in a square and 6 options were presented in 6 other squares around the stimulus word. 3 words related to the stimulus word had to be chosen from among those 6 options. A sample item of this test is presented below. In this item, the word *banana* is the stimulus word and the test-taker should choose *fruit, peel, and yellow* as its associates by drawing a line between them.



In 1998, Read developed another selective WAT task to measure depth of word knowledge in advanced learners of English (Read, 1998). This test also consisted of 40 items containing a stimulus word followed by 8 words. This time, he put the following 8 options for each stimulus, in two boxes. Each box contained 4 words. The test-takers were asked to choose the synonyms of the stimulus word (which was an adjective in this test) from the left box if there was any. The possible collocations of the stimulus adjective were chosen from among the 4 words in the right box. Qian (2002) reported a high correlation coefficient (0.88) between this test and a famous vocabulary size test (VLT) (Nation, 1990). A sample item from this test appears below:

Fertile

special dark private growing business soil egg mind

Qian (1999) also used the second revision of selective test developed by Read (1998) to find the relationship between depth of word knowledge and reading comprehension. He called this revised test DVK (Depth of Vocabulary Knowledge). Each item of a DVK test consisted of a stimulus word followed by 8 words which were presented in two boxes. The model of choosing the options was the same as mentioned above. The only difference was the choice of stimulus words in a way to be familiar to the target sample. One sample item of this test is as follows:

Sudden

beautiful quick surprising thirsty change doctor noise school

The test taker should determine that he should choose the words *quick* and *surprising* from the left column and *change* and *noise* from the right column. He administered this test to 74 adult Chinese and Korean advanced English learners who took a reading comprehension test too. He found that depth of vocabulary knowledge made a unique contribution to the prediction of reading comprehension scores and depth of vocabulary knowledge played a fundamental role in reading comprehension processes and that there was a positive relationship between the learners' depth of vocabulary knowledge and their lexical inferring ability.

Meara and Fitzpatrick (2000) developed a free word association task called Lex 30 to measure productive aspect of deep word knowledge. It contained 30 items which presented the test-taker with a stimulus word and asked him to provide 3 related words to each stimulus. All stimulus words were highly frequent and chosen from Nation's first 1000 wordlist. None of the stimulus words was aimed to elicit a single dominant response. The researchers aimed at including such stimulus words which elicit wide range of response words which stands out of Nation's first1000 word list. So in pilot stage they first ensured that at least half of the native speakers' responses to items were not included in Nations' first 1000 word list.Participants were 46 adult EFL learners with a variety of L1 backgrounds. Eachtestee took a vocabulary Yes/No test too. Then all response words of each testee wereanalyzed with the help of a software, and the number of answers which exceeded Nation's first frequency level of wordlist was considered as his/her productive lexical depth score. The correlation between the results of Lex 30 and receptive Yes/No test was 0.841 ($P < 0.01$). The researchers argued that this test was a more practical and effective way to measure non-native speakers' word knowledge.

Qian and Schedl (2004) tried to evaluate the selective measure of deep word knowledge by comparing its results with the results of a vocabulary section of a new TOEFL test. For this purpose they redeveloped their selective WAT with a team of TOEFL test developers based on the format of word association test developed by Read (1998). They used the stimulus words of TOEFL vocabulary measure so that the DVK (deep vocabulary knowledge test) and TOEFL would have the same set of stimulus words to be compared. The sample item is:

Powerful

potent definite influential supportive repetition price engine position

In the example shown above, the test taker should choose the 4 words of the 8 words provided in the boxes. The synonym of the word *powerful* should be chosen from the left box while its possible collocations should be chosen from the right box. This test and the TOEFL vocabulary test were administered to 207 international students studying English as the second language in a major Canadian university. The purpose of the study was to determine whether DVK could be a useful base for developing an appropriate reading comprehension test.The results showed that the new measure has a similar difficulty level with TOEFL vocabulary test and could predict test-takers' reading performance. The Pearson correlation between these two tests was 0.84 which shows a high overlap between them. The Pearson correlation between DVK and a full standardized TOEFL test was 0.74, which is an acceptable degree of reliability.

Schoonen and Verhallen (2008) investigated the performance of native and non-native speakers of Dutch on selective WAT tasks. They used the same format used in Verhallen and Schoonen (1998). Participants were 795 third and fifth grade children in the age range of 9 to 12, who were studying in 19 different primary schools in Netherlands. The WAT test was in Dutch language. It was shown that younger children possessed less deep lexical knowledge than older children, and L1 monolingual children generally possessed more deep lexical knowledge than L2 bilingual children. They established the concurrent validity of the test by estimating its correlation with a highly reliable definition task and its reliability was estimated using internal consistency alpha Cronboche statistics. They found that native speakers of Dutch have deeper word knowledge in this language compared to their non-native peers and that fifth grade students acted better than 3rd grade students in WAT tasks. They concluded that selective WAT task is appropriate to measure depth of word knowledge in this group of students.

In this study both productive and selective formats of WAT tasks as complementary measures of depth are used to assess depth of word knowledge to provide a rich source of evidence for the purpose of judging about test-takers' deep word knowledge in two groups of elementary and advanced EFL learners. The following research questions were put forth:

1) Does type of WAT formats (productive vs. selective) affect elementary EFL learners' performance on tests of deep word knowledge?

2) Does type of WAT formats (productive vs. selective) affect advanced EFL learners' performance on tests of deep word knowledge?

3. Method

3.1 Design

This study used a quantitative approach to data collection and analysis in which the performance of students on productive and selective formats of test of deep word knowledge was compared applying a within-group comparison design. This comparison was done separately for each of two groups of advanced and elementary EFL learners.

3.2 Participants

Participants of this study were 71 advanced female EFL learners within the age range of 17 to 21 and 82 elementary female EFL learners within the age range of 9 to12 year old. Four intact classes of advanced learners were chosen from Jahade Daneshgahi Language Institute (JDLI) in Urmia, Iran. The overall number of learners in advanced group was 74 female students who were studying in level 14. After administering an initial proficiency test of vocabulary, 3 of them were excluded as outliers. The Passages Series written by Richards and Sandy (2000) (Cambridge University Press) were used as student books for advanced levels.

Four other intact classes of elementary learners were chosen from the same institute. The total number of the elementary learners was 86, all females 4 of whom were excluded as outliers after administering the initial proficiency test. Top Notch Series (Saslow & Ascher, 2006) were used as student books for elementary levels.

3.3 Instrument

This study elicited the necessary data by utilizing 6 different instruments. The first two instruments were two different initial language proficiency tests, one for elementary and the other for advanced group (a standard Key English Test (KET) for elementary level and a CPE (Cambridge Certificate of Proficiency in English) for advanced group). These two tests were administered before the main study for the purpose of homogenizing the participants. Two different selective WAT tasks were developed or adopted and used separately for each group of elementary and advanced learners. Also two different productive WAT tasks were developed or adopted and administered to each group as described below

3.3.1 Instruments Used for Eliciting Data from Elementary Learners

1) KET which is a standard initial proficiency test used at the first session to exclude outliers.

2) A tailor-made WAT which is a selective test format to measure depth of word knowledge in elementary learners (WAT E).

3) A productive format of WAT tasks (similar to Lex 30) to measure depth of word knowledge in elementary learners (Lex E).

WAT E

This test is a selective format of word association task to measure deep vocabulary knowledge in elementary levels. This test was developed by the researchers based on the methods and principles that previous researchers used to develop a selective word association task to measure depth of word knowledge in elementary levels. The original idea comes from Verhallen & Schoonen (2008). They used such a test in their study to measure depth of word knowledge in 3rd and 4th grade students. Before that, depth of word knowledge was measured only in advanced learners or native speakers. For the first time Verhallen and Schoonen tried to measure this construct in elementary Dutch bilinguals. They developed a word association test which presented the young learner with a stimulus word followed by 6 options. The test-taker should chose 3 words from among those 6 words which he/she thinks are related to the stimulus word and each item had 3 points.

In our case, the test was developed in the same manner but the number of items was decreased to 20 because of time limitations. The stimulus words were selected from the students' previous student books used in Jahad institute. The provided options were also checked to be familiar for them. Each item has 3 points with one point given to each correct choice. The total score for this test is 60. Piloting was done in a B1 class of 23 students. In order not to be confused with the WAT task used in advanced classes, the name WAT E is used here to refer to this test. A sample item is shown below:

دانش آموز عزیز لطفا به هر کلمه نگاه کرده و دور3 کلمه که فکرمیکنید به آن مربوط اسث را از بین 6 کلمه ی

داده شده خط بکشید.(فقط دور 3 کلمه خط بکشید)

Bird:

a. fly b. pen c. sing d. music e. chair f. sky

The test-taker who chooses options a, c and f could earn the full credit of this item.

Lex E

This test is a productive format of a WAT task to measure depth of word knowledge in elementary learners. The test is adapted from Istifchi (2005) and piloted in a b1 class of 19 elementary learners in Jahad institute. This free word association test contains 20 items which present the test-taker with a stimulus word and ask them to write 3 words related to that stimulus word. Those words which belong to the word association classification of Verhallen (1994) can earn the test-taker one point with the total score amounting to 60.

Sample item:

دانشآموزعزیزلطفاباتوجهبهنمونه ی زیر 3 کلمهراکهبهمحضدیدنکلماتزیربهذهنتانمیرسد

درمقابلآنهابنویسید.

دانش آموز عزیز لطفا با توجه به نمونه ی زیر 3 کلمه را که به محض دیدن کلمات زیر به ذهنتان میرسد در مقابل آنها بنویسید.

Banana: yellow, fruit, delicious

In this example the student has produced the words *yellow, delicious* and *fruit* for the stimulus word (banana). *Fruit* is the super ordinate association which denotes a paradigmatic relationship and thus gets one point. *Yellow* also indicates a conceptual word relation and gets a point. Although the word delicious is relevant to banana, it does not belong to our criterion set of word associations suggested by Verhallen and Schoonen (1994) and it does not earn the candidate any points.

3.3.2 Data Elicitation Tools Used for Eliciting Data from Advanced Learners

1) CPE initial proficiency test used to establish proficiency and to exclude outliers.

2) WATA, a selective format of WAT (word association tasks) to assess depth of word knowledge in advanced group.

3) Lex 30, a productive format of WAT tasks to measure depth of word knowledge and compare its results with the results gained by the selective format (WATA).

WATA

This test is the selective format of Word Association Tasks (WAT) adopted from Read (1998) which has been wildly used in recent studies on assessment of deep word knowledge. This test is originally named WAT (Word Association Test) by its developer (Read, 1998) but in order to prevent to be confused with the term WAT which stands for Word Association Task here we call it WATA because it was used to elicit data from advanced level.

WATA contains 40 items. Each item consists of a stimulus word and two boxes. The stimulus word is an adjective which is followed by 8 other words which are put in the mentioned two boxes. Each box contains four words. Among four words in the left box the test taker should chose one to three of them which can be the synonym of the stimulus word. Among four words in the right box the test taker should mark one to three of them which can be a collocation for the stimulus adjective. Each item always has just 4 correct answers. In other words, test-takers are allowed to choose 4 words from among the 8 words but these 4 choices are not evenly distributed between two boxes. There are 3 possible situations in choosing the words from the two boxes:

1) The left and right boxes both contain two correct answers.

2) The left box contains one correct choice and the right box contains three correct answers.

3) The left box contains three correct answers and the right box contains 1 correct answer.

Here is an example of a WAT A item:

Powerful

(A) potent (B) definite (C) influential (D) supportive

(E) position (F) engine (G) repetition (H) price

For each correct choice one point is given. So the total score is 160 for 40 items. Although the validity and reliability of this test has been checked and reported in Qian (2002), it was piloted in our study to establish its reliability index in an Iranian context. A reliability coefficient of 0.65 was observed. According to Qian (1998), 35 minutes is enough to attend to all 40 items.

LEX 30

The Lex 30 task is basically a word association task, in which testees are presented with a list of stimulus words, and required to produce responses to these stimuli (Meara & Fitzpatrick, 2000).There is no predetermined set of response target words for the subject to produce, and in this way, Lex 30 resembles a free productive task. Lex 30 consists of 30 items each of which contains a stimulus word and requires test-takers to write the first 3 words they remember by seeing that word. This test was originally designed by Meara and Fitzpatrick (2000) to assess productive vocabulary knowledge. Stimulus words are highly frequent and were chosen from Nation's first 1000 wordlist. There are words which even students of low language proficiency can perceive and produce words for them. None of the stimulus words typically elicit a single special related word. They are chosen to be able to remind the testee of a wide range of associated words. The purpose for which the researcher is using the test determines its model of scoring. For example, to assess productive vocabulary the words produced by testee are checked by software to see the number of words which exceed the first level of1000 words of Nation's frequency wordlist (Nation, 1984). But for assessing depth of word knowledge, the answers can be checked through a word association treasure checklist such as Edinburg Associative Thesaurus or scored according to a special criterion of word association classification.

In this study the answers were scored according to whether they belong to the word association classification suggested by Verhallen and Schoonen (1994). For each answer belonging to this model, one score is given and since just 3 answers for each item is allowed the total score for 30 items would be 90. A reliability of 0.866 ($p < 0.01$) was recorded by Fitzpatrick and Clenton (2010) for this test. A sample item of Lex 30 is:

Produce the first 3 words you remember by seeing this word:

Attack:

The test-taker may produce: war, gun, fear, etc...

For produced word to receive a credit it should belong to one of these categories as suggested by Verhallen (1994):

1) Paradigmatic relationship (subordinates; super ordinates; synonyms, e.g. animal/dog or Plant/flower/rose)

2) Syntagmatic relationship (definitional aspect of a word and possible collocations, e.g. furniture/desk)

3) Partonomic relationship (part-whole relationship) e.g. banana/peel.

4) Conceptual relationship e.g. banana/yellow.

3.4 Procedure

Firstly four intact classes of elementary students were selected from Jahad institute in Urmia, Iran. The overall number of elementary students was 82 young learners attending B1 classes. Data was gathered in 5 sessions as follows:

During the first session, an adapted version of KET was administered and students who got scores which were 2 SD more and less than the mean were excluded. During the second session, the selective format of WAT tasks called WAT E was administered to the subjects. The productive format called Lex E was administered at the 3rd session. The same procedure was followed to elicit data from advanced level students. The initial proficiency test was administered at the first session for the purpose of homogenizing the participants. At the second session the selected format of WAT tasks (WAT A) was administered to all classes. The productive format of WAT task was administered at the 3rd session.

3.5 Data Analyses

Using SPSS software, one paired samples*t*-testwas utilized to compare advanced groups' performance on productive WAT task with their performance on selective WAT task.Another paired *t*-testwas run to compare elementary learners' performance on selective WAT task with their performance on productive WAT task.

4. Results

To answer our research questions, we developed 2 null hypotheses:

H01. There isn't any significant difference between elementary EFL learners' performance on productive WAT tasks and their performance on selective WAT tasks.

H02. There isn't any significant difference between advanced EFL learners' performance on productive WAT tasks and their performance on selective WAT tasks.

After checking for the normality of the groups, a paired samples*t*-test was run in order to check the first null hypothesis. (Table 1)

Table 1. Paired samples test for the elementary level

		Paired Differences			t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean			
Pair 1	Selective test task - Productive test task	4.45	4.27	.56	7.94	81	.00

The results of the paired t-test show that the first null hypothesis is rejected. There is a statistically significant difference between the performance of the elementary participants in the selective and productive tasks: t (81) = 7.94, $p < 0.0$

In other words, the elementary participants' performance on WAT tests is affected by the type of the test task. The following table shows the mean scores for the two tasks while the total score for each task is 60.

Table 2. Descriptive statistics for the elementary level

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Selective test task	50.48	82	6.04	.79
	Productive test task	46.03	82	7.68	1.01

The elementary participants performed better in the selective task rather than the productive one (mean$_{\text{Selective}}$ = 50.48; mean$_{\text{Productive}}$ = 46.03) (Table 2). We can conclude that selective tasks are a better way of measuring young learners' deep word knowledge.

After checking the normality in the advanced group, another paired samples*t*-test was run in order to check the second null hypothesis.

Table 3. Paired samples test for the advanced level

		Paired Differences			t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean			
Pair 1	Selective test task - Productive test task	-4.06	12.30	1.74	-2.34	70	.02

The results of the paired t-test show that the second null hypothesis is rejected. There is a statistically significant difference between the performance of the advanced participants in the selective and productive tasks: t (70) = -2.34, p < 0.05. (Table 3)

In other words, the advanced participants' performance on WAT tests is affected by the type of the test task. The following table shows the mean scores for the two tasks. In order for the two tests to be compared statistically two sets of scores were changed to percentages, a ratio of 100.

Table 4. Descriptive statistics for the advanced level

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Selective test task	52.66	71	19.94	2.82
	Productive test task	56.72	71	17.27	2.44

The advanced participants performed better in the productive task rather than the selective one (mean$_{\text{Selective}}$ = 52.66; mean$_{\text{Productive}}$ = 56.72) (Table 4). This means that productive tasks may bea better measure of depth of word knowledge in this level.

5. Discussion

Our initial aim in conducting this study was to find out that which type of test format is more appropriate for eliciting deep word knowledge from two groups with different language proficiency levels. To be able to develop a valid test for measuring a component of language ability, the test developer should first ensure the reliability of a test. Because in order for a test score to be valid, it needs to be reliable (Bachman, 1996). For achieving a good degree of reliability we need to eliminate those factors which create errors in measurement. In other words, reliability is the extent to which the scores gained in a test is due to the pure language ability which is under measurement. There are factors which affect test scores and test performance other than language ability and these factors are the source of measurement error which reduces reliability of a test (Bachman, 1996). It is a duty for any language tester and test developer to take care of factors which affect test performance other than language ability itself and try to reduce the effect of these factors to decrease measurement error in order to arrive at more reliable and thus valid tests. The investigation of reliability involves both logical analyses and empirical research. We must identify the sources of error and find appropriate ways to reduce their effects to ensure that the scores of a test are very close (if not exactly equal) to test-takers' real ability in the language component under measurement. For example, if we want to measure depth of word knowledge after defining and describing the nature of this component we also need to conduct empirical studies to investigate the other sources which affect the scores in deep word knowledge tests other than the lexical depth itself. These factors can be poor health, fatigue, lack of interest or motivation, cognitive styles, age, sex, positive or negative attitudes toward a test content or format, which create errors in measurement.''In order to identify these sources of error, we need to distinguish the effects of language abilities we want to measure from the effects of other factors, and this is a particularly complex problem. This is partly because of interaction between components of language ability and test method facets which may make it difficult to mark a clear boundary between the ability being measured and the method facet of a given test (Bachman, 1996, p, 161). For example if we want to measure depth of word knowledge in a group of EFL learners, the test format which we choose to elicit the knowledge of this language component may affect the test-takers' performance in a way that their scores are not accurate predictors of their real deep word knowledge. In other words, they may perform better in a productive test format and thus the investigation for finding which format (selective or productive) produces more reliable scores is crucial. An outstanding effort to directly investigate the effect of test format on test performance has been conducted by In'nami and Koizumi (2009) who explored the quality of test performance on multiple-choice and constructed response format of listening and reading tests. They did it by meta-analyzing fifty-six data sources for estimating the mean effect sizes of test format effects. They found that learners' performance is better in multiple-choice format although the test items in both constructed and multiple-choice format were stem-equivalent. In other words, the content of the test and the construct under the measurement was the same but responses were different for each type of test format. It is thus evident that a portion of measurement has been due to error. Our research on the effect of test format on test performance is in line withIn'namiand Koisumi (2009), who found that test format influence test performance. Because there was a significant difference between performance of both elementary and advanced group in two formats of WAT tests.Furthermore the findings of this study regarding the appropriateness of selective WAT tasks for elementary EFL learners are in line with Schoonen and Verhallen (2008) who found selective tasks a good measure of deep word knowledge in 3th and 4th grade EFL learners. The fact that elementary learners perform better in selective tasks may be attributed to their weakness in productive aspect of vocabulary knowledge. Because productive word knowledge is active dimension of lexical repertoire which needs more language exposure and experience, a stage which elementary learners have not reached yet. Advanced learners on the other hand enjoy a high degree of exposure to different uses of words during their long English language courses till level 14.Considering this fact the appropriate test format for advanced group may be the productive task. Although Meara and Fitizpatrick (2000) believes that Lex 30 as a free productive word association task possess a high quality in eliciting productive word knowledge in advanced learners, Cremer et al. (2010) believes that productive WAT tasks are not an appropriate measure of depth of word knowledge and it lacks objectivity because of a large range of free responses. So our findings are homogeneous with the former while in contrast with the latter. Considering the fact that advanced learners' both active and receptive word knowledge is developed in a good degree, they might be comfortable in selective tasks too. Their high scores in the productive task might be attributed to the fact that selective tasks are highly controlled and distractors might include some difficult or non-acquired words which in turn discourage the test-taker to attend a selective item confidently so they miss the credit for that item.

6. Conclusion

This study provides empirical evidence on the effect of test format on test performance. It indicates that by selecting an appropriate test format to assess vocabulary knowledge language teachers can have a deeper understanding of students' progress in expanding their mental lexicon. By adding a section of WAT tasks to the current tools of vocabulary assessment and practices they can even recognize which dimension of vocabulary has not been efficiently worked on and which points need more attention at the side of teachers. Qian (2002) suggests that depth of word knowledge had the highest contribution to lexical inferring success. So attempts for establishing new teaching methods for developing this dimension of word knowledge can be helpful for students' future academic success. Findings of this study can add to teachers' knowledge about the nature of mental lexicon and can be a prompt to create new ways of vocabulary teaching with the use of word association strategies. It is also evident that different learners with different language proficiency demands for different methods of testing even if the construct on which they are tested is the same. While advanced learners perform more effectively in a productive format, the elementary learners act better in a selective test. So findings of this study help to increase teachers' awareness of the needs of their students with different language proficiency and of the importance of the format by which they try to measure their vocabulary knowledge.

For a deeper understanding of what format might be better for each proficiency group of EFL learners, an investigation of participants' attitude toward these two formats can be added to the body of this kinds of studies.

Investigating the relationship between language proficiency and depth of word knowledge is also suggested as further research.

References

Bachman, L. (1996). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Beck, I. L., & McKeown, M. G. (1983). Learning words well-a program to enhance vocabulary and comprehension. *The Reading Teacher*, *36*, 622-625.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman, & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-70). Cambridge: Cambridge University Press.

Christ, T. (2011). Moving Past "Right" or "Wrong" Toward a Continuum of Young Children's Semantic Knowledge. *Journal of Literacy Research*, *43*(2), 130-158. http://dx.doi.org/10.1177/1086296X11403267

Cremer, M., Dingshoff, D., Beer, M. D., & Schoonen, R. (2010). Do word associations assessword knowledge? A comparison of L1 and L2, child and adult word associations. *International Journal of Bilingualism*, *15*(2), 187-204. http://dx.doi.org/10.1177/1367006910381189

Fitizpatrick, T., & Clenton, J. (2010). The challenge of validation: Assessing the performance of a test of productive vocabulary. *Language Testing*, *27*(4), 537-554. http://dx.doi.org/10.1177/0265532209354771

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, *21*, 303-317.

Innami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*(2), 219-244.

Ishtifchi, I. (2005). *Playing with words: A study on word association responses*. Paper presented on 21-23 October 2005 at New Horizons in ELT: An International Conference Inged & Economics and Technology University.

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, *21*, 202-226.

Meara. P., & Fitizpatrick, T. (2000). An improved method of assessing productive vocabulary in an L2. *System*, *28*, 19-30. http://dx.doi.org/S0346-251X(99)00058-5

Meara, P. (1983). Word associations in second language. *Nottingham Linguistics Circular*, *11*, 28-38.

Meara, P. (1997). Towards a new approach to modelling vocabulary acquisition. In N. Schmitt, & M. McCarthy, (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 109-121). Cambridge: Cambridge University Press.

Nagy, W., & Scott, J. (2000). Vocabulary processes. In M. Kamil, P. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 269-284). Mahwah: Lawrence Erlbaum.

Nassaji, H. (2004). The relationship between depth of vocabulary knowledge and L2 learners' lexical inferencing strategy use and success. *Canadian Modern Language Review*, *61*, 107-134.

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

Pearson, P. D., Hiebert, E., & Kamil, M. (2007). Vocabulary assessment: What we know and what we need to learn. *Reading Research Quarterly*, *42*(2), 282-296.

Qian, D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing*, *21*(1), 28-52. http://dx.doi.org/10.1191/0265532204lt273oa

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, *52*(3), 513-536.

Read, J. (2000). *Assessing vocabulary*. Cambridge, UK: Cambridge University Press.

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In B. Harley, & J. H. Hulstijn (Eds.), *Language Learning and Language Teaching. Vocabulary in a second language: Selection, acquisition and testing* (Vol. 10, pp. 209-227). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Read. J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*, 355-371. http://dx.doi.org/10.1177/026553229301000308

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 41-60). Mahwah, NJ: Lawrence Erlbaum.

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

Richards, J. C., & Sandy, C. (2000). *An upper-level multi-skills course Passages*. Cambridge, Cambridge University Press.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge, UK: Cambridge University Press.

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

Schoonen, R., & Verhallen, M. (2008). The assessment of deep word knowledge in young first and second language learners. *Language Testing*, *25*(2), 211.

Schoonen, R., & Verhallen, M. (1998). *Aspects of Vocabulary Knowledge and Reading Performance*. Paper presented at the Annual Meeting of the American Educational Research Association.

Snow, C. E., & Locke, J. (2001). Quality of phonological representations of lexical items. *Applied Linguistics*, *22*, 283-477.

Verhallen, M. (1994). *Lexicale vaardigheden van Turkse en Nederlandse kinderen. Een vergelijkend onderzoek naar betekenistoekenning* (Doctoral dissertation with an English summary). University of Amsterdam, the Netherlands.

Verhallen, M., & Schoonen, R. (1998). Lexical knowledge in L1 and L2 of third and fifth graders. *Applied Linguistics*, *19*(4), 452-470.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics*, *22*, 217-234.

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review*, *53*, 13-40.

Zevara, A. (2005). Models of lexical knowledge assessment of second language learners of English at higher levels of language proficiency. *System*, *33*, 547-562.

Copyrights