https://doi.org/10.17323/jle.2022.13422

Comparing Two Measures of L2 Depth of Vocabulary Knowledge Using their Relationship with Vocabulary Size

Ali Dabbagh [1], **Mostafa Janebi Enayat** [2]

[1] Gonbad Kavous University, Iran
[2] University of Maragheh, Iran

ABSTRACT

Background. This study compared two tests of second language (L2) depth of vocabulary knowledge, namely the word association test (WAT) and vocabulary knowledge scale (VKS), with respect to their associations with vocabulary size. The same relationships were further examined separately for the five word-frequency bands of the vocabulary size test.

Methods. 115 Iranian English as a Foreign Language (EFL) learners who were native speakers of Persian took the WAT, VKS, and Vocabulary Levels Test (VLT). The selected participants were undergraduates who ranged from freshmen to junior and were both male (n=47) and female (n=68) students.

Results. The outcomes of multiple linear regression analyses indicated that: (a) while both measures of vocabulary depth were predictive of the VLT, the WAT had a higher association with the dependent variable; (b) both the WAT and VKS were predictive of the high-frequency vocabulary, with the relationships being more significant for the WAT; (c) the WAT could significantly predict the mid-frequency vocabulary, whereas the VKS had no significant contribution; and (d) while the VKS was significantly associated with the low-frequency vocabulary, the WAT had no significant contribution to the prediction of this level.

Implications. The implications of the findings are interpreted with reference to the suitability of both the WAT and VKS depending on the type of input, expected response, and desired frequency of the target words.

KEYWORDS

vocabulary size, vocabulary depth, Word Association Test (WAT), Vocabulary Knowledge Scale (VKS), Vocabulary Levels Test (VLT)

INTRODUCTION

Vocabulary knowledge has been recognized as one of the most significant components of language learning without which no meaning can be conveyed and understood (Dabbagh & Janebi Enayat, 2019; Duong, 2022; Janebi Enayat & Babaii, 2018; Mathews, 2018; Read, 2004; Roche & Harrington, 2013; Schmitt, 2010; Schmitt, 2014; Schmitt & Schmitt, 2014). Uchihara and Clenton (2022), for instance, found that spoken vocabulary knowledge is significantly correlated with second language (L2) speaking proficiency. Researching vocabulary involves dealing with a multidimensional construct as the nature of this knowledge is perplexing and entails various aspects of form, meaning, and use, each of which encompasses sub-components (Laufer et al., 2004; Nation, 1990; Schmitt, 2014). To grapple with such complexity, a variety of descriptive frameworks have been suggested to systematically categorize the construct of vocabulary knowledge, the most oft-cited of which is the classification of *size* and *depth* (Haastrup & Henriksen 2000; Henriksen, 1999; Qian, 1998; Read 1993; Schmitt, 1999), with the former pertaining to the number of words L2 learners know during a particular stage of the learning process (Nation, 2001) and the latter relating to the quality of word knowledge or how well the L2 learners know a single lexical item (Read, 2000). Depth of vocabulary knowledge, therefore, embodies not only the diction-

Citation: Dabbagh, A., & Janebi Enayat, M. (2022). Comparing Two Measures of L2 Depth of Vocabulary Knowledge Using their Relationship with Vocabulary Size. *Journal of Language and Education, 8*(3), 53-65. https://jle.hse.ru/article/view/13422

Correspondence:
Mostafa Janebi Enayat
m.enayat@maragheh.ac.ir

Received: November 25, 2021
Accepted: June 08, 2022
Published: September 30, 2022

ary definition of a word, but also its semantic network which includes, but is not limited to, paradigmatic and syntagmatic lexical relations (Schoonen & Verhallen, 2008), which refer to the linear relations between two words that could appear in the same sentence (e.g., research-conduct, research-observation, and research-laboratory) and hierarchical relations (e.g., research-science, research-experimental), respectively.

The introduction of these two dimensions of vocabulary knowledge led to the development of some reliable and valid tests to measure them. Vocabulary size, in particular, has attracted more attention in L2 vocabulary research (David, 2008) due to its critical and substantial contribution to effective language use (Alharthi, 2020; Dabbagh & Janebi Enayat, 2019; Derakhshan & Janebi Enayat, 2020; Nguyen & Nation, 2011; Uchihara & Clenton, 2020). Various instruments have been, therefore, developed to examine the L2 learners' vocabulary size. Vocabulary Levels Test (VLT) is perhaps the most widely used measure of vocabulary size (Webb & Sasao, 2013) which was first designed by Nation (1983) and later revised and validated by Schmitt et al. (2001). The test is built upon the five word-frequency levels of 2,000, 3,000, 5,000, and 10,000 that comprise 120 high- and low-frequency target words. Another measure of vocabulary size is the Vocabulary Size Test (VST) designed by Nation and Beglar (2007) and validated by Beglar (2010) that evaluates L2 vocabulary size using fourteen 1,000-word-frequency levels that include 140 lexical items. This test has addressed more word-frequency bands which have made it more comprehensive (Elgort, 2013) and suitable to measure the progress of vocabulary size over time (Beglar, 2010). Another measure of vocabulary size is the Yes/No test format designed by Meara (1992) which, compared to the VLT, was found to be a less effective test (Cameron, 2002). A more recent and modified version of the VLT, known as the New Vocabulary Levels Test (NVLT), has been designed by Webb et al. (2017). The NVLT tests the L2 vocabulary size using the first five 1,000-word-frequency bands.

The tests of vocabulary size, such as the VLT, VST, and NVLT are based on the word-frequency bands. These tests start with high-frequency vocabulary like knowledge of the first 1,000-word-frequency level and end with low-frequency vocabulary, such as the 10,000-word-frequency band in the VLT and the 14,000-word-frequency level in the bilingual and monolingual versions of the VST (Janebi Enayat et al., 2018). Schmitt and Schmitt (2014) reassessed the boundaries and proposed another category. They argued that high-frequency English vocabulary should contain the most frequent 3,000 word families. Additionally, they proposed that the low-frequency vocabulary should be lowered to 9,000-word frequency level and beyond. The authors labelled the vocabulary between high-frequency (3,000) and low frequency (9,000) as the "mid-frequency" vocabulary.

Depth of vocabulary knowledge, however, has gained less attention in language testing as it entails a range of word relations, making it difficult to offer a unified definition for this dimension of word knowledge (Milton, 2009; Schmitt, 2014). In fact, compared to the number of tests developed and validated for measuring vocabulary size, "less progress has been made, both in defining depth as a construct and in developing tests for practical use" (Read, 2007, p. 105). The first is Vocabulary Knowledge Scale (VKS), designed by Pribakht and Wesche (1993) and Wesche and Paribakht (1996), which assesses different stages of vocabulary knowledge ranging from no familiarity with the word to the ability to use it accurately in sentences. However, the instrument which could find its way in almost all of the previous studies on depth of vocabulary was Word Association Test (WAT). Developed and validated first by John Read in 1993, WAT assesses depth of vocabulary knowledge through asking learners to choose only four out of eight responses which may be, in one way or another, related to the cue word. As Read (1993) stated, "it is assumed that learners with a deeper knowledge of the word will be better able to pick the associates (which should represent different aspects of the meaning of the word) than those whose knowledge is more superficial" (p. 395).

Despite the surge of interest in using WAT as a valid measure of depth of vocabulary knowledge, recent studies have revealed that this test might partially score vocabulary size, as well. Previous studies showed significant correlations between scores resulted from WAT and VLT, a measure of vocabulary size (Akbarian, 2010; Huang, 2006; Janebi Enayat et al., 2018; Noro, 2002; Schmitt, 2014). This high interrelationship between the two constructs is in line with the argument proposed by Meara and Wolter (2004) that the two aspects of depth and size are not separate from each other and improvement in vocabulary size results in the development of vocabulary depth as well. The current study has attempted to find the more suitable test of vocabulary depth for university English as a Foreign Language (EFL) learners using this interrelationship between the two aspects of depth and size of vocabulary knowledge, using WAT and VKS as measures of vocabulary depth and VLT and its frequency bands as a measure for vocabulary breadth.

LITERATURE REVIEW

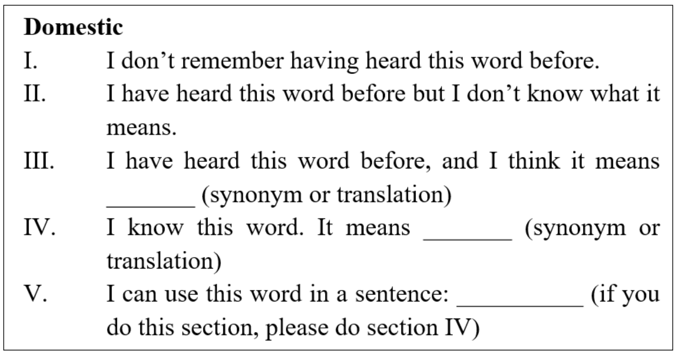
In this section, literature on two main measures of vocabulary depth, i.e., WAT and VKS, and their limitations are briefly reviewed. Then, the previous research on probing the interaction between vocabulary breadth and depth is succinctly reviewed. Highlighting Meara and Wolter's (2004) model as the theoretical background utilized in the present study, the section ends with introducing the research gap and formulating the research questions.

Measures of Vocabulary Depth

The depth dimension of vocabulary knowledge has been measured using a couple of tests, but, compared to the number of measures developed for size aspect, less attempt has been made to both define this dimension as a construct and develop tests to measure it (Read, 2007). Read (2000) classified the different approaches to measure depth of vocabulary knowledge into two main groups: developmental and dimensional. In the former, a scale of measurement is used to describe the stages in vocabulary acquisition. To that end, Paribakht and Wesche (1993) and Wesche and Paribakht (1996) designed VKS. First designed to assess English vocabulary learning in language programs at the University of Ottawa, Canada, this scale measures the different levels of lexical knowledge of particular words being learned in a comprehension-based ESL classroom. This scale is a self-report measure which learners are asked to specify their degree of understanding of individual words on a scale of 1-5. The first three categories of this scale deal with conceptual familiarity with the cue word (from no familiarity to the ability to provide a synonym) and the last two categories involve assessing the productive knowledge of the prompt words by asking to compose a response (category IV: I know this word. It means _____ and category V: I can use this word in a sentence, as follows) (see Figure 1).

Figure 1

Sample of VKS item (Taken from Wesche & Paribakht, 1996)



Domestic

I. I don't remember having heard this word before.

II. I have heard this word before but I don't know what it means.

III. I have heard this word before, and I think it means _______ (synonym or translation)

IV. I know this word. It means _______ (synonym or translation)

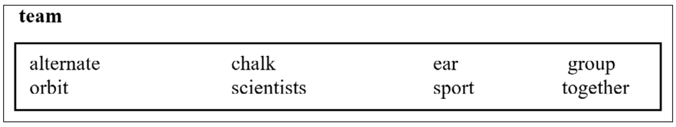
V. I can use this word in a sentence: __________ (if you do this section, please do section IV)

However, as Qian (1998) argued, VKS assesses only one meaning of the prompt word coupled with its actual use and ignores measuring multiple meanings or associations. Henriksen (1999) further confirmed this argument and noted that VKS only assesses the receptivity or productivity of the target words with no measurement of their different aspects. In addition, Schmitt (2010) listed the following limitations for this scale: 1. the first two stages of the scale are unverified; 2. the underlying knowledge construct are inconsistent, jumping from form-meaning (categories I to IV) to production in context (category V); 3. the intervals between the categories are not consistent; 4. the metalinguistic judgement in categories II (I *think* I know the word) and III (I *know* the word) can be confusing for some learners since they are better at judging what they can do with the words; and, more importantly, 5. the simple sentences examinees write in category V cannot clearly show their productive knowledge of the target word. As Webb (2013) mentioned, in VKS, "it is possible [for test takers] to use a word correctly in a sentence without knowing its meaning" (p.3). In this regard, Zhong (2016) suggested to adapt the test in a way to reach the minimum possible chance for test takers to produce 'neutral' sentences like 'It is beautiful' or 'He is calm'.

The dimensional approach, on the other hand, tries to describe the mastery of various components of different words and considers the mastery of lexical networks of an individual word as important (Read, 1993). To assess such an aspect, WAT was designed and further revised by Read (1993, 1998) which assesses the depth of individual vocabulary knowledge through word association and the relationships between the words in the mental lexicon. This was a developed format of his previous attempt to measure depth of vocabulary through interview procedure in which the learners were asked to pronounce the words, provide an explanation, identify the domain, provide word associations, and suggest other forms of the word (Read, 1998). The first version of the test designed in 1993 includes eight options for each target word of which four were associated with the target word paradigmatically (synonym), syntagmatically (collocation), and analytically (component) (see Figure 2).

Figure 2

The first version of WAT in 1993



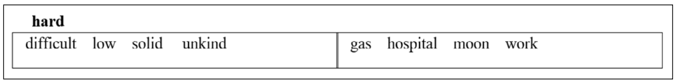
team

alternate	chalk	ear	group
orbit	scientists	sport	together

The 1998 version uses two boxes with eight words in each for 40 target words, all of which are adjectives. The examinees should select only four words associated with the target word from the two boxes (see Figure 3). The words in the left box are paradigmatically related to the target word and the ones in the right box are syntagmatically related. To reduce the guessing effect, the patterns of students' responses differ such that three format are possible: two words from the right box and two from the left one; three from the right and one from the left; or three from the left and one from the right.

Figure 3

The 1998 version of WAT



hard

difficult low solid unkind	gas hospital moon work

The merit of this test format is in its ability to tap different instances of meaning, collocation, and formulaic language (Schmitt, 2010). Schmitt *et al*. (2011) reported that WAT could be regarded as an appropriate measure of depth of vocabulary since "it is tapping into learners' uncertainty about collocational combinations" (p.118).

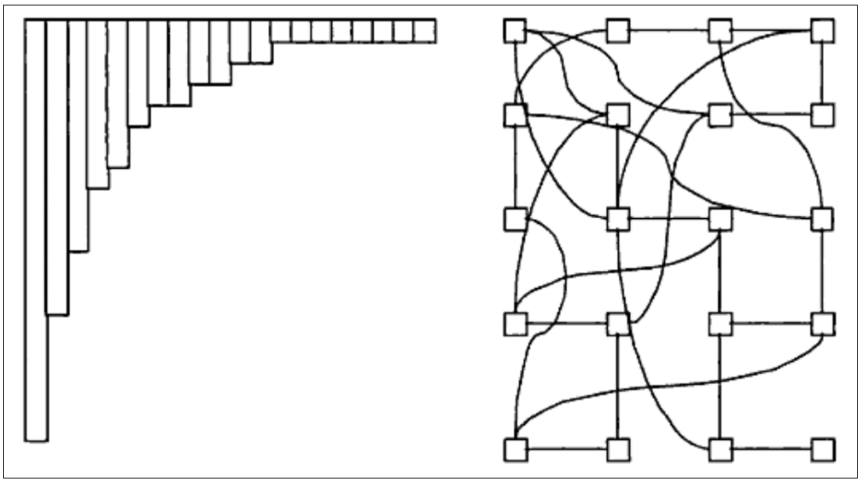
Despite its wide use in measuring depth of vocabulary knowledge (e.g., Akbarian, 2010; Atai & Dabbagh, 2010; Dabbagh, 2016; Dabbagh & Janebi Enayat, 2019; Janebi Enayat & Babaii, 2018; Janebi Enayat & Derakhshan, 2021; Janebi Enayat et al., 2018; Nassaji 2006, Qian 1999, 2002; Schoonen & Verhallen, 2008, among others), WAT is regarded as a challenging measure of depth of vocabulary knowledge for advanced learners at university level (Greidanus et al., 2005; Greidanus & Neinhuis, 2001; Zhang & Koda, 2017). In addition, different scholars refer to the shortcomings of WAT as a measure of vocabulary depth from various viewpoints. Webb (2013) pointed out that although WAT measures three different aspects of vocabulary depth, namely concept and referents, form and meaning, and collocation, it does not provide separate scores for each of these aspects and it is plausible that two test takers who are actually distinct in their depth of vocabulary dimensions receive the same score without being distinguished in terms of what depth of vocabulary aspect was known by each. Akbarian (2010) highlighted that due to the identification of nouns to be collocated with the adjectives given in the test as target words, the test addresses knowledge of adjectives directly and nouns rather indirectly. Also, adverbs are indirectly focused on in WAT since almost all adverbs are related to their corresponding adjectives (Ishii, 2005). However, measuring depth of knowledge of verbs is taken for granted and not included in the test. In addition, as Milton (2009) and Read (1993, 1998) asserted, WAT is susceptible to guessing due to its receptive multiple-choice format which can threaten the validity of the test. Test takers can easily choose some of the given words on random which can make the score interpretation problematic since scores may not provide a true estimate of the test takers' depth of vocabulary knowledge. As Schmitt *et al.* (2011) asserted in their validation study of WAT, the guessing effect can mostly happen for scores 0-2 and not for scores 3-4 for each item. More specifically, they found that split scores – where test takers achieve 1, 2, or 3 out of the maximum 4 for each item – mostly resulted from no knowledge or partial knowledge of the target word and consequently no clear interpretation can be reached upon for these scores. They also relate guessing in WAT items to its tendency to overestimate the test takers' actual knowledge of the target words and raise the question of whether test takers are successful in guessing even if they have no knowledge of the target words.

The Interconnection between Size and Depth as a Possible Yardstick

Even though being distinct in terms of measurement instrument, the depth and size of vocabulary have been found to be so much inter-related. Nurweni and Read (1999), in a study on the vocabulary knowledge of first-year students in an Indonesian university concluded that the tests of size (word translation test) and depth of vocabulary (WAT) correlated highly with each other (*r* = .62). Qian (1999) explored this issue and found a significant correlation of .78 between the VLT and WAT scores among 44 Korean and 33 Chinese speakers. Henriksen (1999) further argued that "an understanding of the relations among the items is a prerequisite for a more precise understanding of each individual item" (p. 313). This interconnection between the two dimensions of size and depth of vocabulary was supported by Meara and Wolter (2004) who believed that "vocabulary size is not a feature of individual words: rather it is a characteristic of the test taker's entire vocabulary" (p. 87). These two scholars proposed two different models for the interconnection between size and depth of vocabulary (see Figure 4). While in the first one (the left hand diagram) vocabulary size and depth are not intrinsically interrelated and adding more lexical items does not develop the whole lexicon, in the second model (the right hand diagram), an increase in vocabulary size could develop the lexical network (vocabulary depth) as well. This relationship was also approved by Ishii and Schmitt (2009) who contended the two aspects are interconnected such that one dimension would be incomplete without the other.

Figure 4

Two ways of looking at the relation between vocabulary size and depth. From "V-Links: beyond vocabulary depth," by P. Meara and B. Wolter, in D. Albrechtsen, K. Haastrup, and B. Henriksen (Eds.), Writing and vocabulary in foreign language acquisition (p. 89), 2004, Museum Tusculanum. Press



Many studies have reported the interconnection between these two dimensions of vocabulary size and depth, as measured by the VLT and WAT, respectively (e.g., Akbarian, 2010; Gyllstad, 2007; Huang, 2006; Janebi Enayat et al., 2018; Qian, 2002; D. Zhang, 2012). This interconnection, however, does not mean that the test takers' scores on the VLT could show both size and depth of vocabulary knowledge because tests of receptive vocabulary size intend to measure form/ meaning recognition knowledge, and not vocabulary depth which is assessed using word associations tasks. Put it more simply, the relationship between these two dimensions could possibly mean that they are related to the same construct and, therefore, should not be seen as separate aspects (Vermeer, 2001). Another interpretation is that tests of depth of vocabulary knowledge, such as WAT, are not actually tests of vocabulary depth; they are rather size tests "masquerading as depth tests" (Akbarian, 2010, p. 400). This claim was also backed by Milton (2009) who asserted that the associative format to measure depth of vocabulary is not successful in measuring this vocabulary construct for the main reason that this format is incapable of tapping into the quality of association the test takers make.

The correlation between size and depth of vocabulary knowledge has been reported to be unclear for lower and higher frequency words. While there seems to be little difference between these two dimensions for higher frequency words, a gap has been reported between these aspects of vocabulary for lower frequency words (Schmitt, 2014). Shimamoto (2000), Noro (2002), and Henriksen (2008) for instance, found the relationship to be weaker for learners who had larger vocabularies and higher language proficiency.

In the present study, the proposed model of Meara and Wolter (2004) and the interconnection between the two dimensions of vocabulary size and depth were utilized as a yardstick to identify the most suitable test of vocabulary depth. Additionally, the nature of this relationship was probed for higher and lower word-frequency levels of the vocabulary size test. Therefore, the following research questions were formulated in this study: (1) Which measure of vocabulary depth has the highest predictive ability for L2 vocabulary size? (2) How is the predictive ability of the two measures of vocabulary depth in L2 vocabulary size different for high and low word-frequency bands?

METHODS

In this section, the demographic information of the participants as well as the instruments used for data collection are explained. The steps followed for data collection and analysis are also described.

Participants

A sample of 115 intermediate EFL undergraduate learners, who were all native speakers of Persian, was selected based on the results of the quick Oxford Placement Test (2004) out of 234 Iranian undergraduate students of English Language Teaching and English Language and Literature. Accordingly, the participants who scored between 30 and 47 in the test, i.e., B1 and B2 according to the Common European Framework of Reference (CEFR), were selected. The selected participants ranged from freshmen to junior who were both male (n=47) and female (n=68) students with the age range of 18 to 25. The reason for selecting this sample is that based on the nature of the study, participants should have a good mental lexicon in terms of quality and quantity of word knowledge and an acceptable command of English.

Instruments

Oxford Quick Placement Test (OQPT, 2004)

To homogenize the participants in terms of the proficiency level, this test was administered and the ones with intermediate level of English language proficiency were selected. The test, which was developed by Oxford University Press and University of Cambridge Local Examinations Syndicate, consists of 60 multiple-choice items to which participants were to answer in 30 minutes. According to the OQPT scoring system, the participants who scored between 26 and 45 were determined as intermediate language level. A high validity and a reliability close to .90 was reported by Geranpayeh (2003) for this version of the test.

WAT-Test of Dimensional Aspect of Vocabulary Depth

Developed by Read (1993), WAT measures the depth of vocabulary knowledge of the participants. The test is a list of 40 prompt words each of which consists of one stimulus word, which is an adjective, followed by a list of eight words in two boxes of four words. The left and right boxes consist of the synonymous words and collocations of the stimulus words, respectively. The participants should choose four words that are related to the prompt word semantically. The four related words have been selected to represent three semantic relations, namely paradigmatic, syntagmatic and analytic (Read, 1993). Read (1995) reported its reliability (KR-20, N=94) as .93 and Nassaji (2006) and Qian (2002) found its split half reliability to be .89.

VKS- Test of Developmental Aspect of Vocabulary Depth

Developed originally by Paribakht and Wesche (1993), VKS was used to find out the participants' self-perceived level of developmental aspect of depth of vocabulary knowledge. Participants should indicate their level of knowledge about the target words on a Likert scale ranging from total unfamiliarity to the ability to use the words in context. The instrument enjoys a high reliability estimate of .89 for content words and .82 for discourse connectives as reported by Paribakht and Wesche (1997).

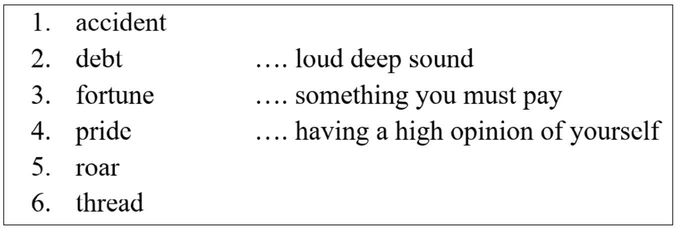
As VKS is a tool which in theory can be used with any set of words and since the aim of the current study is to compare VKS and WAT, the same prompt words in the latter test were utilized as the cue words for the former.

VLT-Test of Vocabulary Size

Designed by Nation (1983) as a measure of breadth of vocabulary, this test "provides a profile of a learner's vocabulary" (p.58) in terms of levels of frequency (2000-, 3000-, 5000-, and 10,000-word-frequency levels) with large samples of words from different frequency levels. In other words, "[the test scores] obtained from VLT were treated as the variable of size of vocabulary knowledge" (Akbarian, 2010, emphasis added). The test has been validated and revised by many scholars since its first format (e.g., Ishii & Schmitt, 2009; Schmitt et al., 2001; Xing & Fulcher, 2007). Version 2 of this test, which was revised and validated by Schmitt *et al.* (2001), is employed in this study. In this version, the participants were given 10 groups of words in each frequency level. Each group consists of 6 cue words that should be matched with 3 definitions (see Figure 5). The test has been reported as reliable with a Cronbach alpha of .96 (Akbarian, 2008) and .81 (Schmitt et al., 2001).

Figure 5.

A VLT sample item (Taken from Schmitt et al., 2001)



1. accident	
2. debt	…. loud deep sound
3. fortune	…. something you must pay
4. pride	…. having a high opinion of yourself
5. roar	
6. thread	

Although Nation and Beglar's (2007) Vocabulary Size Test (VST) has been claimed to be a more comprehensive measure of breadth of vocabulary than the VLT, the present study utilized the latter for the reason that the four-option multiple-choice format of VST is subject to guessing effect (Gyllstad *et al.,* 2015), which may lead to the overestimation of test scores over and above the six-option matching format in VLT (Stewart, 2014; Stewart & White, 2011). Moreover, VLT is the widely used measure of breadth of vocabulary among researchers (e.g., Abdullah et al., 2013; Akbarian, 2010; Alavi & Akbarian, 2012; Baba, 2009; Dabbagh, 2016; Dabbagh & Janebi Enayat, 2019; Janebi Enayat & Derakhshan, 2021; Janebi Enayat et al., 2018; Qian, 2002; Webb & Sasao, 2013; Zhang & Anual, 2008).

Procedure and Data Analysis

First, the Oxford Quick Placement Test was administered to determine participants' proficiency level and select the intermediate ones. Then, to measure the participants' depth and size of vocabulary knowledge, WAT, VKS and VLT were administered with a one-week time interval for each test to prevent sensitization of students to the purpose of the research and control the testing effect. While administering the WAT, the participants were encouraged to give as many answers as they could, even if they would not be sure whether the given answers were correct or not (Read, 1993). As for the VLT, the participants were required not to follow the guessing strategy for the words they did not know, but they were suggested to find the answer if they thought they might know it. The time allotted for each test was 30 to 45 minutes. The WAT, VKS, and VLT papers of the participants were scored following the criteria established by Nassaji (2006), Wesche and Paribakht (1996), and Schmitt *et al.* (2001), respectively. Multiple linear regression analyses were run using SPSS version 23.0 to find the contribution of WAT and VKS to VLT and the extent that the high and low word-frequency bands were predicted by the two tests of vocabulary depth.

RESULTS

Descriptive and Reliability Statistics

Table 1 represents a general profile of the descriptive statistics of the participants' scores on the WAT, VKS, VLT, and the four word-frequency bands of the VLT. As the data analyzed below shows (see Table 1), the participants' scores on the three administered vocabulary tests and the sub-tests of the VLT enjoyed appropriate Cronbach's alpha reliability estimate, which means these tests enjoy high reliability for the sample of the present study.

Before running multiple regression analyses, the correlations among the variables were calculated. The results of Shapiro-Wilk test indicated that except for the scores on the VKS and VLT, the scores on the other WAT and the four sub-tests of the VLT were not normally distributed ($p > .05$). Spearman correlation coefficients were calculated for the sets of scores, the results of which are provided in Table 2. It shows that the correlations among all the variables were significant ($p < .05$) and the correlations between the VKS and WAT, as the predictor variables were also significant ($p < .05$). However, multicollinearity, i.e., correlation of independent variables in a regression model (Field, 2009), was not a concern as the tolerance values were less than 0.40 and the variance inflation factors (VIFs) were less than 2.5 (Field, 2009).

Table 1

Descriptive statistics for the participants' scores on each test and sub-test.

Test	MPS	Min.	Max.	Mean	SD	α
WAT	100	18	65	46.99	9.84	.83
VKS	200	75	182	138.22	20.30	.86
VLT	120	33	85	64.14	11.15	.89
VLT 2,000	30	20	30	27.49	2.38	.74
VLT 3,000	30	10	29	21.94	4.43	.76
VLT 5,000	30	1	24	12.75	5.23	.79
VLT 10,000	30	0	6	1.94	1.49	.81

Note: N = 115. MPS = Maximum possible score; SD = standard deviation; α = Cronbach's alpha.

Table 2

Spearman correlation coefficients among the vocabulary depth and size tests and sub-tests

Test	WAT	VKS	VLT	VLT 2K	VLT 3K	VLT 5K	VLT 10K
WAT	-						
VKS	.313**	-					
VLT	.433**	.430**	-				
VLT 2K	.396**	.363**	.671**	-			
VLT 3K	.454**	.397**	.832**	.522**	-		
VLT 5K	.337**	.283**	.895**	.467**	.586**	-	
VLT 10K	.231*	.422**	.666**	.428**	.375**	.640**	-

*Note: N = 115. K = 1,000. **p < .01 **p < .05*

Predictive Ability of WAT and VKS in VLT

To answer the first research question, the contribution of the participants' WAT and VKS scores to VLT scores was examined through multiple linear regression analysis (using the stepwise method). The results, as shown in Table 3, revealed that two models emerged for this association. The first model in which only the WAT was entered as the predictor variable could explain about 23% of the variance in the VLT (*F (1,113) = 33.565, p < .001,* R^2 = .229). The second model where both WAT and VKS were entered as the explanatory variables could explain 29% of the VLT performance (F (2,112) = 23.052, p < .001, R^2= .292). In other words, the addition of the VKS scores could provide an additional 6% of the predictive power which was a significant change (p < .01). This shows that WAT had a more significant association with the test of vocabulary size compared to the VKS test. Put it simply, the receptive format of vocabulary depth was more predictive of the scores on the test of vocabulary size than the productive format.

The standardized beta weights also reaffirmed the strength of the association between the scores on the WAT and VLT in the first (β = .479, t = 5.794, p < .001) and second (β = .355, t = 4.005, p < .001) models. The VKS, however, made a less contribution to the prediction of the VLT scores (β = .279, t = 3.146, p < .01).

Predictive Ability of WAT and VKS in High and Low Frequency Vocabulary of VLT

The second research question of the current study investigated the extent that the WAT and VKS scores could predict the high and low word-frequency bands of the VLT. A series of multiple linear regressions (using the stepwise method) were run for this purpose. The results (see Table 4) indicated that, for the 2,000-word-frequency band of the VLT, two models emerged. In the first model, only the WAT was entered as the predictor variable which could explain 15.5% of the variance in this sub-test of the VLT (*F (1,113) = 20.781, p < .001,* R^2 = .155). The second model in which both WAT and VKS were entered as the predictor variables could explain 19.5% of this word-frequency band of the VLT (F (2,112) = 13.580, p < .001, R^2= .195). The addition of the VKS scores could, therefore, add 4% to the predictive power which was a significant change (p < .05). Similar results were found for the 3,000-word-frequency band as two models emerged for this dependent variable in the first of which only the WAT was entered capable of explaining 24% of the variance in the scores of this sub-test (F (1,113) = 35.722, p < .001, R^2= .240). In the second model where both WAT and VKS were entered as the explanatory variables, the predictive power was 31.5% (F (2,112) = 25.752, p < .001, R^2= .315), indicating that the additional variance explained by the insertion of the VKS was about 7% which was statistically significant (p < .01). As for the 5,000-word-frequency band, one model emerged in which WAT was the only predictor variable capable of explaining 14% of the variance in the scores obtained on this sub-test of the VLT (F (1,113) =8.575, p < .001, R^2= .141). In contrast, in the one model appeared for the 10,000-word-frequency band of the VLT, it was the VKS scores which could significantly provide a similar prediction for the dependent variable (F (1,113) = 19.075, p < .001, R^2= .144). The results, therefore, indicated that the WAT was more associated with the high- and mid-frequency vocabulary size which are measured through the 2,000-, 3,000-

Table 3

Multiple regression analyses for vocabulary depth measures in vocabulary size.

	R	*R²*	*ΔR²*	Unstandardized		Standardized
				B	SE *B*	β
Model 1	.479	.229***				
Constant				38.664	4.493	
WAT				.542	.094	.479***
Model 2	.540	.292***	.063**			
Constant				24.044	6.349	
WAT				.402	.100	.355***
VKS				.153	.049	.279**

***p < 0.01, ***p < 0.001.*

Table 4

Multiple regression analyses for vocabulary depth measures in word-frequency levels of the VLT

Dependent	Predictor	*R*	*R²*	*ΔR²*	Unstandardized		Standardized
					B	SE *B*	β
VLT 2K	Model 1	.394	.155***				
	Constant				23.009	1.005	
	WAT				.095	.021	.394***
	Model 2	.442	.195***	.040*			
	Constant				20.516	1.447	
	WAT				.072	.023	.296**
	VKS				.026	.011	.223*
VLT 3K	Model 1	.490	.240***				
	Constant				11.568	1.774	
	WAT				.221	.037	.490***
	Model 2	.561	.315***	.075**			
	Constant				5.211	2.483	
	WAT				.160	.039	.355***
	VKS				.067	.019	.305**
VLT 5K	Model 1	.376	.141***				
	Constant				3.372	2.224	
	WAT				.200	.046	.376***
VLT 10K	Model 1	.380	.144***				
	Constant				-1.911	.893	
	VKS				.028	.006	.380***

**p < 0.05, **p < 0.01, ***p < 0.001.*

(high-frequency words), and 5,000-word-frequency vocabulary (mid-frequency vocabulary), while the VKS was more linked with the 10,000-word-frequency band that relates to the low-frequency vocabulary size·

Appraisal of the standardized beta further confirmed the significant associations between the WAT scores and the 2,000-word-frequency band (β = .394, t = 4.559, p < .001), the 3,000-word-frequency level (β = .490, t = 5.977, p < .001), and the 5,000-word-frequency level (β = .376, t = 4.310, p < .001). The links between the VKS scores and the 2,000-word-frequency band (β = .223, t = 2.354, p < .05) as well as the 3,000-word-frequency level (β = .305, t = 3.497, p < .01) were comparatively less significant. In contrast, while the WAT performance was the only variable associated with the 5,000-word-frequency level (β = .376, t = 4.310, p < .001), the VKS was the only format which could be linked with the 10,000-word-frequency band (β = .380, t = 4.368, p < .001).

DISCUSSION

The current study was an attempt to identify the most suitable measure of vocabulary against the yardstick of associations with VLT, a measure of vocabulary size. The findings of multiple linear regression analyses for the scores of 115 EFL students indicated that the WAT was more significantly associated with the VLT scores, particularly the high- and mid-frequency bands. The VKS, however, had a comparatively weaker contribution to the prediction of the VLT scores, but its prediction of the low-frequency band of this test was unique.

The findings indicated that the interconnection between size and depth, as two aspects of vocabulary knowledge, was strong, as measured through the WAT and VLT, supporting previous studies (Akbarian, 2010; Gyllstad, 2007; Henriksen, 2008; Milton, 2009; Zareva, 2005). For instance, Akbarian (2010) used regression analysis and reported that WAT could predict the variance in the VLT. This study also found that the links between the higher frequency words of the VLT and WAT were stronger than the 10,000-word-frequency band. This could somehow support Schmitt's (2014) conclusion that for higher levels of vocabulary size "there is often little difference between size and a variety of depth measures" while this association is weak for lower frequency bands of the VLT where "there is often a gap between size and depth, as depth measures lag behind the measures of size" (p. 941). Noro (2002) and Henriksen (2008) further reported a less significant correlation between the VLT and WAT for lower frequency words. The strong association between the two tests could be justified with reference to the findings of Meara and Wolter (2004) who reported that an increase in vocabulary size could lead to an increase in vocabulary depth, particularly for lower levels of language proficiency. In a more recent study, Janebi Enayat and Amirian (2020) found that the VLT and WAT are significantly correlated, particularly for lower-intermediate students. The association, however, was not high for advanced L2 learners. Dabbagh and Janebi Enayat (2019) further found high correlations between the size and depth aspects of vocabulary knowledge, as measured through the VLT and WAT, respectively. This means that the two dimensions could jointly contribute to the overall L2 language proficiency. For instance, Janebi Enayat and Derakhshan (2021) investigated the contribution of vocabulary size and depth to L2 speaking ability using the VLT and WAT and found that the two aspects could jointly predict the L2 oral proficiency.

The results further indicated that the prediction of the VLT was mainly made by the WAT while the VKS contributed to the prediction of the VLT scores less significantly. This could be due to the different task format of the WAT, which employs matching items, while the VKS uses a scale that indicates knowledge subjectively. The objective matching format of the WAT is more compatible with the matching type of the VLT which both reduce the guessing effect (Stewart, 2014). Therefore, the students' score on the WAT could be a more precise indication of their depth of vocabulary knowledge than the VKS which is more subjective. The findings also showed a lower power of VKS than that of WAT in predicting VLT. This finding implies that WAT can be regarded as a measure of depth of vocabulary that is more influenced by the size dimension of vocabulary knowledge, and hence, according to Meara and Wolter's (2004) model, it might be regarded as a better measure of depth of vocabulary in comparison with the VKS.

The results for the second research question revealed that, surprisingly, while the WAT was more predictive for the high- and mid-frequency vocabulary (Schmitt & Schmitt, 2014), for the 10,000-word-frequency band of the VLT, the VKS was the only predictor variable. This can be further discussed in that the partial receptive/productive nature of the VKS can better picture knowledge of less frequent vocabulary compared to the WAT which is only receptive. As it was mentioned previously rather implicitly, the first three columns of VKS measure receptive aspects of depth of vocabulary knowledge and the other two columns focus on the productive aspect. This special feature of VKS makes it more suitable to measure depth of vocabulary both receptively and productively. On the contrary, WAT is mainly a receptive measure of vocabulary depth dealing with making associations among the given words. The difference between receptivity and productivity of these two depth of vocabulary tests can be regarded as the cause of their distinction in regression analysis results. However, there is no doubt that low-frequency vocabulary would be recognized to a larger extent than being mastered productively because these words are supposed to be the most difficult ones in the test. This could be justified further by taking into account the fact that productivity is not always about making complex meaningful sentences and even writing a simple synonym would make a test productive. Additionally, the fourth scale (see Figure 1), which is identified as the productive part of the scale, asks the students to provide a synonym or simply a translation for the target word. Consequently, the students could just write a translation for the target word which could be easier for the students than encountering the target word in a receptive test which requires knowing some other words in the list of options. What is more, due to the format of the WAT where the students must choose four words from a list of eight words for a target word (see the instruments sections for further information on the patterns of responses), the four responses for the target word are dependent on each other, which could make this "receptive" format more difficult than the fourth scale of the VKS where only a translation would suffice to inform the examiner that the student may have a partial knowledge of the target word. As a result, the probability of knowing a low-frequency word could be more on the VKS than the WAT. This provides empirical support for Read's (2004) proposal calling for distinguishing among different aspects of depth of vocabulary with different measures.

Taking the overall results into account, it can be claimed that although WAT was shown to be more predictive as a measure of vocabulary size, and hence a better measure for depth of vocabulary than VKS in this regard. Each of these tests should be used depending on the purpose of measurement, i.e., whether to measure receptive or productive aspects of depth of vocabulary. Moreover, for tapping less frequent aspects of vocabulary depth, the VKS would be a more suitable option as it has the expected response of only providing a synonym or a translation for the target word.

CONCLUSION

The current study compared WAT and VKS in order to find the most appropriate measure of vocabulary depth via comparing their power to predict VLT scores, as a measure of vocabulary size. It can be concluded that although the WAT scores explain the variance in the VLT scores to a larger extent and could be, therefore, considered a more suitable test of vocabulary depth when we consider the association of size and depth as a yardstick, the VKS should also be seen as a more subjective test of vocabulary depth that could tap into the more productive aspect of this dimension of vocabulary knowledge. The results shed light on the difference between WAT and VKS reporting a low correlation between the two which signifies that they cannot be used for research and instruction purposes interchangeably. Rather they should be used for the purposes which correspond to the nature of their test item structure. In other words, vocabulary researchers can use VKS when they are exploring the role of depth of vocabulary in speaking and writing performance, as productive skills, especially if the focus of the investigation is on less frequent words. Also, WAT can be used in probing the association between reading and/or listening comprehension, as receptive skills, and vocabulary depth. With this specification of the use of measures of depth of vocabulary knowledge, more precise results might be achieved in future vocabulary studies.

The results and conclusion of the present study need to be interpreted with caution as there were some limitations which lead to some suggestions for further research. First of all, similar to previous quantitative studies on WAT, VKS, and VLT, this investigation was based on correlational design and quantitative data. Further qualitative can deepen our understanding of the learners' perceptions and processes in answering items of these tests. Second, this study used the VLT, which is not a comprehensive test of vocabulary size. Future studies can be conducted using different measures of vocabulary size, such as Peabody Picture Vocabulary Test (PPVT) and Vocabulary Size Test (VST) to find about the overlap of depth of vocabulary knowledge with other aspects of vocabulary size. Third, this study focused on one language proficiency level to control the effect of this variable and homogenize the students. The interaction among WAT, VKS, and VLT can be assessed benefiting EFL learners from different proficiency levels.

DECLARATION OF COMPETING INTEREST

None declared.

REFERENCES

Abdullah, K. I., Puteh, F., Azizan, A. R., Hamdan, N. N., & Saude, S. (2013). Validation of a controlled productive Vocabulary Levels Test below the 2000-word level. *System, 41(2), 352–364.* https://doi.org/10.1016/j.system.2013.03.005

Akbarian, I. (2008). *The role of vocabulary knowledge in predicting performance on reading comprehension item types [Unpublished doctoral dissertation]. University of Tehran, Iran.*

Akbarian, I. (2010). The relationship between vocabulary size and depth for ESP/EAP learners. *System, 38*(3), 391–401. https://doi.org/10.1016/j.system.2010.06.013

Alavi, S. M., & Akbarian. I. (2012). The role of vocabulary size in predicting performance on TOEFL reading item types. *System, 40*(3), 376–385. https://doi.org/10.1016/j.system.2012.07.002

Alharthi, T. (2020). Investigating the relationship between vocabulary knowledge and FL speaking performance. *International Journal of English Linguistics, 10(1), 37–46.* https://doi.org/10.5539/ijel.v10n1p37

Atai, M. R. & Dabbagh, A. (2010). Investigating vocabulary depth and semantic set in EFL learners' vocabulary use in writing. *TELL Journal, 4(2), 27–49.*

Baba, K. (2009). Aspects of lexical proficiency in writing summaries in a foreign language. *Journal of Second Language Writing, 18*(3), 191–208. https://doi.org/10.1016/j.jslw.2009.05.003

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27(1), 101–118.* https://doi.org/10.1177%2F0265532209340194

Cameron, L. (2002). Measuring vocabulary size in English as an additional language. *Language Teaching Research, 6(2), 145–173.* https://doi.org/10.1191%2F1362168802lr103oa

Dabbagh, A. (2016). The predictive role of vocabulary knowledge in listening comprehension: depth or breadth? *International Journal of English Language and Translation Studies, 4(3), 1–13.*

Dabbagh, A., & Janebi Enayat, M. (2019). The role of vocabulary breadth and depth in predicting second language descriptive writing performance. *Language Learning Journal, 47(5), 575–590.* https://doi.org/10.1080/09571736.2017.1335765

David, A. (2008). Vocabulary breadth in French L2 learners. *The Language Learning Journal, 36(2), 167–180.* https://doi.org/10.1080/09571730802389991

Derakhshan, A., & Janebi Enayat, M. (2020). High- and mid-frequency vocabulary size as predictors of Iranian university EFL students' speaking performance. *Iranian Journal of English for Academic Purposes, 9(3), 1–13.*

Duong, T. M. (2022). Insights into ESP Vocabulary Learning Strategies Used by Vietnamese Tertiary Students. *Journal of Language and Education, 8*(1), 38–49. https://doi.org/10.17323/jle.2022.10924

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing, 30(2), 253–272.* https://doi.org/10.1177%2F0265532212459028

Field, A. (2009). *Discovering statistics using SPSS (3*rd ed.). Sage Publications.

Geranpayeh, A. (2003). A quick review of the English Quick Placement Test. *UCLES Research Notes, 12, 8–10.*

Greidanus, T., & Nienhuis, L. (2001). Testing the quality of word knowledge in a second language by means of word associations: Types of distracters and types of associations. *Modern Language Journal, 85*(4), 567–577. https://doi.org/10.1111/0026-7902.00126

Greidanus, T., Beck, B., & Wakely, R. (2005). Testing the development of French word knowledge by advanced Dutch and English-speaking learners and native speakers. *Modern Language Journal, 89*(2)*, 221–233.* https://doi.org/10.1111/j.1540-4781.2005.00276.x

Gyllstad, H. (2007). *Testing English collocations. Media-Tryck, Lund University.*

Gyllstad, H., Vilkaite, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple choice formats: Issues with guessing and sampling rates. *ITL International Journal for Applied Linguistics, 166*(2), 278–306. https://doi.org/10.1075/itl.166.2.04gy1

Haastrup, K., & Henriksen. B. (2000). Vocabulary acquisition: acquiring depth of knowledge through network building. *International Journal of Applied Linguistics, 10*(2), 221–240. https://doi.org/10.1111/j.1473-4192.2000.tb00149.x

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition, 21*(2), 303–317. https://doi.org/10.1017/S0272263199002089

Henriksen, B. (2008). Declarative lexical knowledge. In D. Albrechtsen, K. Haastrup, & B. Henriksen (Eds.), *Vocabulary and writing in a first and second language* (pp. 22–66). *Palgrave Macmillan.*

Huang, H.-F. (2006). *Breadth and depth of English vocabulary knowledge: Which really matters in the academic reading performance of Chinese university students?* [Unpublished master's thesis]. McGill University.

Ishii, T. (2005). *Diagnostic tests of vocabulary knowledge for Japanese learners of English* [Unpublished doctoral dissertation]. University of Nottingham.

Ishii, T., & N. Schmitt. (2009). *Developing an integrated diagnostic test of vocabulary size and depth. RELC Journal, 40*(5), 5–22. https://doi.org/10.1177%2F0033688208101452

Janebi Enayat, M., & Amirian, S. M. R. (2020). The relationship between vocabulary size and depth for Iranian EFL learners at different language proficiency levels. *Iranian Journal of Language Teaching Research, 8(2), 97–114.* https://dx.doi.org/10.30466/ijltr.2020.120891

Janebi Enayat, M., & Babaii, E. (2018). Reliable predictors of reduced redundancy test performance: The interaction between lexical bonds and test takers' depth and breadth of vocabulary knowledge. *Language Testing, 35(1), 121–144.* https://doi.org/10.1177/0265532216683223

Janebi Enayat, M., & Derakhshan, A. (2021). Vocabulary size and depth as predictors of second language speaking ability. *System, 99, 1–15.* https://doi.org/10.1016/j.system.2021.102521

Janebi Enayat, M., Amirian, S. M. R., Zareian, G., & Ghaniabadi, S. (2018). Reliable measure of written receptive vocabulary size: Using the L2 depth of vocabulary knowledge as a yardstick. *Sage Open, 8(1), 1–15.* https://doi.org/10.1177/2158244017752221

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, *21*(2), 202–226. https://doi.org/10.1191%2F0265532204lt277oa

Mathews, J. (2018). Vocabulary for listening: Emerging evidence for high and mid-frequency vocabulary knowledge. *System*, *72*, 23–36. https://doi.org/10.1016/j.system.2017.10.005

Meara, P., & Wolter, B. (2004). V-Links: beyond vocabulary depth. In D. Albrechtsen, K. Haastrup, & B. Henriksen (Eds.), *Writing and vocabulary in foreign language acquisition* (pp. 85–96). Museum Tusculanum Press.

Milton, J. (2009*). Measuring second language vocabulary acquisition. Multilingual Matters.*

Nassaji, H. (2006). The relationship between depth of vocabulary knowledge and L2 learners' lexical inferencing strategy use and success. *The Modern Language Journal, 90*(3), *387–40.* https://doi.org/10.1111/j.1540-4781.2006.00431.x

Nation, I.S.P. (1983). Testing and teaching vocabulary. *Guidelines, 5*, 12–25.

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

Nation, I.S.P. (2001). *Learning vocabulary in another language.* Cambridge University Press.

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Japan Association for Language Teaching, 31*(7), *9–12.*

Nguyen, L.T.C., & Nation, I.S.P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. RELC Journal, 42(1), 86–99. https://doi.org/10.1177%2F0033688210390264

Noro, T. (2002). The roles of depth and breadth of vocabulary knowledge in reading comprehension in EFL. *ARELE: Annual Review of English Language Education in Japan*, *13*, 71–80. https://doi.org/10.20581/arele.16.0_141

Nurweni, A. & Read, J. (1999). The English vocabulary knowledge of Indonesian university students. *English for Specific Purposes, 18*(2), 161–175. https://doi.org/10.1016/S0889-4906(98)00005-2

Paribakht, T. S., & Wesche, M. (1993). Reading comprehension and second language development in a comprehension-based ESL program. *TESL Canada Journal, 11*(1)*,* 9–29. https://doi.org/10.18806/tesl.v11i1.623

Paribakht, T.S., & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady & T. Huckin *(Eds.), Second language vocabulary acquisition: A rationale for pedagogy (pp.174–200). Cambridge University Press.*

Qian, D. (1998). *Depth of vocabulary knowledge: assessing its role in adults' reading comprehension in English as a second language.* [Unpublished doctoral dissertation]. University of Toronto, Canada.

Qian, D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *The Canadian Modern Language Review, 56*(2), 283–307. https://doi.org/10.3138/cmlr.56.2.282

Qian, D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: an assessment perspective*. Language Learning, 52*(3), 513–536*.* https://doi.org/10.1111/1467-9922.00193

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10(3), 355–371.* https://doi.org/10.1177%2F026553229301000308

Read, J. (1995). Validating the word association format as a measure of depth of vocabulary knowledge. *Paper presented at the 17th Language Testing Colloquium, Long Beach, CA.*

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

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

Read, J. (2004). Plumbing the depth: how should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language, (pp. 209–227)*. Benjamins.

Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies, 7(*2), 105–125. https://doi.org/10.6018/ijes.7.2.49021

Roche, T., & Harrington, M. (2013). Recognition vocabulary knowledge as a predictor of academic performance in an English as a foreign language setting. *Language Testing in Asia*, *3*(12), 1–13. https://doi.org/10.1186/2229-0443-3-12

Schmitt, N. (1999). The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge. *Language Testing, 16(2), 189–216.* https://doi.org/10.1177%2F026553229901600204

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual. Palgrave Macmillan.*

Schmitt, N. (2014). Size and depth of vocabulary knowledge: what the research shows. *Language Learning, 64*(4), 913–951. https://doi.org/10.1111/lang.12077

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, 47(4), 484–503.* https://doi.org/10.1017/S0261444812000018

Schmitt, N., Ching Ng, J. W., & Garras, J. (2011). The word associates format: Validation evidence. *Language Testing, 28(1), 105–126.* https://doi.org/10.1177%2F0265532210373605

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behavior of two new versions of the vocabulary levels test. *Language Testing, 18(1), 55–88.* https://doi.org/10.1177%2F026553220101800103

Schoonen, R., & Verhallen, M. (2008). The assessment of deep word knowledge in young first and second language learners. *Language Testing, 25(2), 211–236.* https://doi.org/10.1177%2F0265532207086782

Shimamoto, T. (2000). An analysis of receptive vocabulary knowledge: Depth versus breadth. *The Japan-British Association for English Teaching, 4, 69–80.*

*S*tewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly, 11*(3), 271–282. https://doi.org/10.1080/15434303.2014.922977

Stewart, J., & White, D. A. (2011). Estimating guessing effects on the Vocabulary Levels Test for differing degrees of word knowledge. *TESOL Quarterly, 45(2), 370–380.* https://doi.org/10.5054/tq.2011.254523

Uchihara, T., & Clenton, J. (2020). Investigating the role of vocabulary size in second language speaking ability. Language Teaching Research, 24(4), 540–556. https://doi.org/10.1177/1362168818799371

Uchihara, T., & Clenton, J. (2022). The role of spoken vocabulary knowledge in second language speaking proficiency. *The Language Learning Journal. Advance online publication.* https://doi.org/10.1080/09571736.2022.2080856

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics, 22*(2), 217–234. https://doi.org/10.1017/S0142716401002041

Webb, S. (2013). Depth of vocabulary knowledge. In C.A. Chapelle (Ed.), *The encyclopedia of applied linguistics (pp. 1–8). Blackwell Publishing Ltd.*

Webb, S. A., & Sasao, Y. (2013). New directions in vocabulary testing. *Language Testing, 44*(3), 263–277. https://doi.org/10.1177%2F0033688213500582

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ILT International Journal of Applied Linguistics, 168(1), 33–69.* https://doi.org/10.1075/itl.168.1.02web

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *The Canadian Modern Language Review, 53*(1), 13–40. https://doi.org/10.3138/cmlr.53.1.13

Xing, P., & Fulcher, G. (2007). Reliability assessment for two versions of Vocabulary Levels Tests. *System, 35(2), 182–191.* https://doi.org/10.1016/j.system.2006.12.009

Zareva, A. (2005). Models of lexical knowledge assessment of second language learners of English at higher levels of language proficiency. *System, 33(4), 547–562.* https://doi.org/10.1016/j.system.2005.03.005

Zhang, B., & Annual, S. B. (2008). The role of vocabulary in reading comprehension: the case of secondary school students learning English in Singapore. *RELC Journal, 39(1), 51–76.* https://doi.org/10.1177%2F0033688208091140

Zhang, D. (2012). Vocabulary and grammatical knowledge in L2 reading comprehension: A structural equation modeling study. *Modern Language Journal, 96, 554–571.* https://doi.org/10.1111/j.1540-4781.2012.01398.x

Zhang, D., & Koda, K. (2017). Assessing L2 vocabulary depth with word associates format tests: Issues, findings, and suggestions. *Asian-Pacific Journal of Second and Foreign Language Education, 2(1), 1–30.* https://doi.org/10.1186/s40862-017-0024-0

Zhong, H. F. (2016). The relationship between receptive and productive vocabulary knowledge: a perspective from vocabulary use in sentence writing. *The Language Learning Journal, 46(4), 1–17.* https://doi.org/10.1080/09571736.2015.1127403