T.C

BURSA ULUDAG UNIVERSITY

INSTITUTE OF EDUCATIONAL SCIENCES

DEPARTMENT OF ENGLISH LANGUAGE EDUCATION

A COMPARISON OF THREE DIFFERENT VOCABULARY SIZE TESTS FOR TESTING LEXICAL COMPETENCE IN AN EFL CONTEXT

M.A. THESIS

Sezen AKSU BALAMUR

BURSA
2019

T.C.

BURSA ULUDAĞ ÜNİVERSİTESİ

EĞİTİM BİLİMLERİ ENSTİTÜSÜ

YABANCI DİLLER EĞİTİMİ ANA BİLİM DALI

İNGİLİZ DİLİ EĞİTİMİ BİLİM DALI

İNGİLİZCE'NİN YABANCI DİL OLARAK ÖĞRETİLDİĞİ ORTAMLARDA SÖZCÜK YETERLİLİĞİNİN ÖLÇÜLMESİNDE ÜÇ FARKLI SÖZCÜK BİLGİSİ TESTİNİN KIYASI

YÜKSEK LİSANS TEZİ

Sezen AKSU BALAMUR

Danışman

Doç. Dr. Levent UZUN

BURSA
2019

BİLİMSEL ETİĞE UYGUNLUK

Bu çalışmadaki tüm bilgilerin akademik ve etik kurallara uygun bir şekilde elde edildiğini beyan ederim.

Sezen AKSU BALAMUR

12.06.2019

EĞİTİM BİLİMLER ENSTİTÜSÜ

YÜKSEK LİSANS İNTİHAL YAZILIM RAPORU

ULUDAĞ ÜNİVERSİTESİ

EĞİTİM BİLİMLER ENSTİTÜSÜ

YABANCI DİLLER EĞİTİMİ ANA BİLİM DALI BAŞKANLIĞI'NA

Tez Başlığı / Konusu: İngilizce'nin Yabancı Dil Olarak Öğretildiği Ortamlarda Sözcük Yeterliliğinin Ölçülmesinde Üç Farklı Sözcük Bilgisi Testinin Kıyası

Yukarıda başlığı gösterilen tez çalışmamın a) Kapak sayfası, b) Giriş, c) Ana bölümler ve d) Sonuç kısımlarından oluşan toplam 188 sayfalık kısmına ilişkin, 08.04.2019 tarihinde şahsım tarafından Turnitin adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan özgünlük raporuna göre, tezimin benzerlik oranı % 15'tir.

Uygulanan filtrelemeler:

1- Kaynakça hariç
2- Alıntılar hariç/dahil
3- 5 kelimeden daha az örtüşme içeren metin kısımları hariç

Uludağ Üniversitesi Eğitim Bilimleri Enstitüsü Tez Çalışması Özgünlük Raporu Alınması ve Kullanılması Uygulama Esasları'nı inceledim ve bu Uygulama Esasları'nda belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini saygılarımla arz ederim.

08.04.2019

Adı Soyadı : Sezen AKSU BALAMUR

Öğrenci No : 801410016

Ana Bilim Dalı: Yabancı Diller Eğitimi Ana Bilim Dalı

Programı: İngiliz Dili Eğitimi

Statüsü : Yüksek Lisans

Doç. Dr. Levent UZUN

08.04.2019

YÖNERGEYE UYGUNLUK ONAYI

"A Comparison of Three Different Vocabulary Size Tests for Testing Lexical Competence in an EFL Context" adlı Yüksek Lisans tezi, Bursa Uludağ Üniversitesi Eğitim Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırlanmıştır.

Tezi Hazırlayan	Danışman
Sezen AKSU BALAMUR	Doç. Dr. Levent UZUN

Yabancı Diller Eğitimi ABD Başkanı

Prof. Dr. Ayla GÖKMEN

T.C.

BURSA ULUDAĞ ÜNİVERSİTESİ

EĞİTİM BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜ'NE,

Yabancı Diller Eğitimi Ana Bilim Dalı'nda 801410016 numara ile kayıtlı Sezen AKSU BALAMUR'un hazırladığı "İngilizce'nin Yabancı Dil Olarak Öğretildiği Ortamlarda Sözcük Yeterliliğinin Ölçülmesinde Üç Farklı Sözcük Bilgisi Testinin Kıyası" başlıklı Yüksek Lisans çalışması ile ilgili tez savunma sınavı, 12.06.2019 günü 10.30-12.00 saatleri arasında yapılmış, sorulan sorulara alınan cevaplar sonunda adayın tezinin (**başarılı /** ~~**başarısız**~~) olduğuna (**oy birliği /** ~~**oy çokluğu**~~) ile karar verilmiştir.

Üye (Tez Danışmanı ve Sınav Komisyonu Başkanı)
Doç. Dr. Levent UZUN
Bursa Uludağ Üniversitesi

Üye
Dr. Öğr. Üyesi İsmet ÖZTÜRK
Bursa Uludağ Üniversitesi

Üye
Dr. Öğr. Üyesi İpek KURU GÖNEN
Anadolu Üniversitesi

Abstract

Yazar	: Sezen AKSU BALAMUR
Author	: Sezen AKSU BALAMUR
University	: Bursa Uludag University
Field	: Foreign Language Education
Branch	: English Language Education
Degree Awarded	: M.A.
Page Number	: XVII+171
Degree Date	: 25/06/2019
Thesis	: A Comparison of Three Different Vocabulary Size Tests for Testing Lexical Competence in an EFL Context
Supervisor	: Assoc. Prof. Dr. Levent UZUN

A COMPARISON OF THREE DIFFERENT VOCABULARY SIZE TESTS FOR TESTING LEXICAL COMPETENCE IN AN EFL CONTEXT

In this study, three receptive vocabulary size tests in similar designs to the Yes/No Test (Y/N Test) (Meara, 1992), Vocabulary Size Test (VST) (Nation & Beglar, 2007a, 2007b) and Vocabulary Levels Test (VLT) (Nation, 2001) and covering the first 5K words in bilingual format were used to measure English receptive vocabulary knowledge of university preparatory class students. These tests, though being in different formats, are assumed to be measuring receptive knowledge of vocabulary at the same level and treated as equivalent. However, there does not exist a study which uses these tests all together and handles the matter whether they measure receptive word knowledge at the same level, so this study aimed to contribute to the field by filling this gap. Towards this aim, the study questioned whether the three tests, as assumed, can estimate overall receptive vocabulary size offering similar statistical figures a) in different proficiency levels, b) across test sections and c) in different frequency bands for different proficiency levels groups. Beside this, how well the tests correlate with each other and which one correlates best with the participants' university preparatory class exit scores and yields consistent results were investigated.

The data were collected from elementary, pre-intermediate and intermediate level 581 students studying at two different universities in Turkey and statistically analysed through the SPSS 22.0 packet program. The tests were administered in two sessions.

The findings put forth that in general, the Y/N Test provided the lowest scores in almost all frequency and proficiency based comparisons. The VLT and VST, though, presented diverse scores for different proficiency groups and at different frequency bands. According to the statistical analyses, all mean scores were in line with the participants' level of proficiency, which meant higher level students did better in all frequency levels, and the scores declined accordingly as participants proceeded to lower frequency bands. In addition, the highest correlation was found between the VST and VLT, and lastly, the most correlated test with the students' university preparatory class exit scores was the VLT. All these results are important in terms of creating awareness in EFL teachers about the fact that there are different tests which they can substitute for institutional exams to measure students' vocabulary knowledge, and that they can use these tests in different contexts according to students' individual differences.

Keywords: EFL, English Receptive Size of Vocabulary, English Vocabulary Tests, form-meaning link

Özet

Yazar	: Sezen AKSU BALAMUR
Üniversite	: Bursa Uludağ Üniversitesi
Ana Bilim Dalı	: Yabancı Diller Eğitimi Ana Bilim Dalı
Bilim Dalı	: İngiliz Dili Eğitimi Bilim Dalı
Tezin Niteliği	: Yüksek Lisans Tezi
Sayfa Sayısı	: XVII+171
Mezuniyet Tarihi	: 25/06/2019
Tez	: İngilizce'nin Yabancı Dil Olarak Öğretildiği Ortamlarda Sözcük Yeterliliğinin Ölçülmesinde Üç Farklı Sözcük Bilgisi Testinin Kıyası
Danışmanı	: Doç. Dr. Levent UZUN

İNGİLİZCE'NİN YABANCI DİL OLARAK ÖĞRETİLDİĞİ ORTAMLARDA SÖZCÜK YETERLİLİĞİNİN ÖLÇÜLMESİNDE ÜÇ FARKLI SÖZCÜK BİLGİSİ TESTİNİN KIYASI

Bu çalışmada, üniversite hazırlık sınıfı öğrencilerinin İngilizce pasif sözcük bilgilerini ölçmek amacıyla dizayn açısından Yes/No Test (Y/N) (Meara, 1992), Vocabulary Size Test (VST) (Nation & Beglar, 2007a, 2007b) ve Vocabulary Levels Test (Nation, 2001) (VLT)'e benzerlik gösteren ve iki dilli formatta ilk 5000 kelimeyi kapsayan üç pasif sözcük bilgisi testi kullanılmıştır. Bu testler farklı formatta olmalarına rağmen pasif sözcük bilgisini aynı derecede ölçüyor varsayılmakta ve eşdeğer kabul edilmektedir. Fakat bu üç farklı kelime testinin hepsini birlikte kullanarak bunların pasif sözcük bilgisini aynı seviyede ölçüp ölçmediğini ele alan bir çalışma bulunmamaktadır; dolayısıyla bu çalışma bu boşluğu doldurarak alana katkıda bulunmayı amaçlamıştır. Bu amaç doğrultusunda, çalışma, üç testin öne sürüldüğü üzere genel pasif sözcük dağarcığını benzer istatistiksel rakamlar sunarak a)

İngilizce yeterlilikleri farklı gruplarda, b) testlerin farklı bölümlerinde ve c) İngilizce yeterlilikleri farklı gruplar için farklı sıklık aralıklarında ölçüp ölçemediğini sorgulamıştır. Bunun yanı sıra, testlerin birbirleriyle ne derece ilişkili oldukları ve hangisinin katılımcıların üniversite hazırlık sınıfı geçme notlarıyla en fazla ilişkili gösterdiği ve tutarlı sonuçlar verdiği de incelenmiştir.

Veriler Türkiye'de iki farklı üniversitede okuyan temel, orta düzey öncesi ve orta seviyelerdeki 581 öğrenciden toplanmış ve SPSS 22.0 paket program aracılığıyla istatistiksel analizi yapılmıştır. Testler iki oturumda uygulanmıştır.

Bulgular, genel olarak Y/N testin neredeyse tüm sıklık ve yeterlik bazlı karşılaştırmalarda en düşük skorları verdiğini göstermiştir. VLT ve VST ise farklı yeterlik grupları için farklı sıklık bentlerinde değişik skorlar ortaya koymuştur. İstatistiksel analizlere göre, tüm ortalama değerler katılımcıların yeterlik seviyeleriyle uyumlu çıkmıştır ki bu da üst seviye öğrencilerin tüm sıklık aralıklarında daha başarılı oldukları ve skorların katılımcılar daha düşük frekans bentlerine doğru ilerledikçe düştüğü anlamına gelmiştir. Ayrıca, en yüksek korelasyon VST ve VLT arasında bulunmuş ve son olarak da öğrencilerin üniversite hazırlık sınıfı geçme notlarıyla en fazla ilintili test VLT olmuştur. Tüm bu sonuçlar İngilizce'yi yabancı dil olarak öğreten öğretmenlerde öğrencilerin sözcük bilgilerini ölçmede kurumsal sınavların yerine kullanabilecekleri farklı testlerin var olduğu ve bu testleri öğrencilerin bireysel farklılıklarına göre farklı bağlamlarda kullanabilecekleri gerçeği konusunda farkındalık oluşturulması açısından önemlidir.

Anahtar Sözcükler: Yabancı dil eğitimi, İngilizce pasif sözcük dağarcığı, İngilizce sözcük testleri, biçim-anlam ilişkisi

Acknowledgements

I dedicate this thesis to my beloved daughters,

Eylül & Elif BALAMUR

Many people helped me in this thesis journey, so I would love to offer thanks to each one of them. First and foremost, special thanks go to my former supervisor Asst. Prof. Dr. Meral ÖZTÜRK, whose sustained support and constant guidance I appreciated throughout this thesis work. She has inspired and motivated me greatly to conduct this study, and her remarkable expertise in the field as well as her helpful suggestions helped me a lot to follow the path to the completion of this thesis. She was always available and willing to provide advice and constructive comments by e-mail and in person throughout the whole process.

I am also endlessly thankful to Assoc. Prof. Dr. Levent UZUN, my present supervisor, for his valuable guidence and generous contributions. It has been a great honor to complete this thesis under the professional mentoring of him.

In addition, I would like to thank the administration of Bursa Uludağ University School of Foreign Languages, especially the late director Prof. Dr. Şeref KARA. I also owe a lot to the associate directors Mehmet Doğan and Nihan Özuslu, for their permission to conduct this study in the institution and personal encouragement.

I feel greatly indebted to Murat BAYRAK, the director of Bursa Tecnical University, for making it possible to carry out my study in the institution. A big thank you goes to him and his students. In addition, I wish to sincerely thank each and every one of the students, who voluntarily took the tests and eagerly participated in this study.

A group of my colleagues have also assisted me in a variety of ways from the beginning to the end of my thesis work, and for this I am very grateful: Berrin Manga Çetinavcı, Emine-Yıldırım, Sedef Akgül, Bahar Adıyaman, Zülkade T. Gazi, Paul James Brosnihan, Cicely Rose Hazell, Brittany Landorf and Melanie Claire Proteau Blake for their invaluable suggestions and assistance in preparing the bilingual versions of the vocabulary tests and for their careful reading, constructive and critical comments and useful feedback on the revised test items and distractors; Yasemin AKARSU and Ebru ERKAL for inviting their students to participate in the study; and Zeynep SAKA for her support and help with the pilot study and data collection process.

I wish to sincerely express my thanks and appreciation to a former colleague Sıdıka ÖZEMRE for arranging and coordinating test administration at Bursa Technical University. In general, acknowledgements are gratefully extended to all my colleagues for their most invaluable contributions in various ways to the completion of this work.

On a personal note, I will be indebted to my husband Turgay forever, for his valuable support, unflagging encouragement, and much needed distractions from tiring thesis work; my daughters Eylül and Elif, who sacrificed a huge amount of family time to ensure that I can work in an uninterrupted and opportune way and whose love was my ultimate source of motivation, and my father in Heaven, Kemal AKSU, for the spiritual support he offered me. Each of you has share in this accomplishment, and I feel indebted to you all.

This thesis would not come to fruition provided that I was devoid of the strong support of many other people too numerous to mention here. Last but not least, I wish to give my special thanks to all of them.

Table of Contents

Page Number

List of Tables

Table *Page*

List of Figures

Figure *Page*

List of Abbreviations

VLT: Vocabulary Levels Test

VST: Vocabulary Size Test

Y/N: Yes/No Test

L1: First Language

L2: Second Language

ESL: English as a Second Language

EFL: English as a Foreign Language

ELT: English Language Teaching

CEFR: the Common European Framework of Reference for Languages

BNC: the British National Corpus

COCA: the Corpus of Contemporary American English

AWL: the Academic Word List

UWL: the University Word List

1K: the most frequent 1,000 words in the BNC

2K: the second most frequent 1,000 words in the BNC

FA: False Alarm

MC: Multiple Choice

Chapter I

Introduction

In the literature on second language research, it has been well established that knowledge of vocabulary has an essential role in the process of mastering a second/foreign language (Alderson 2005; Milton 2009). Nobody learns vocabulary for its own sake even in their first language. It is no different in ESL and EFL contexts, in which there are demands for the utilization of knowledge. Since vocabulary is considered as an essential component of language, its relation to different language abilities and its importance in second-language (L2) knowledge in terms of academic achievement have always been a concern in the research field (Willis & Ohashi, 2012). Also, vocabulary knowledge occupies a vital place in vocabulary testing as it is among the major variables used to evaluate a learner's performance in different vocabulary tests (Enayat & Amirian, 2016). This explains the reason why both teachers and researchers show a special interest in vocabulary assessment, especially by focusing on its two major aspects: receptive vocabulary size which supplies information on the students' receptive abilities (reading & listening), and productive vocabulary size which indicates the degree to which their productive skills (speaking & writing) can extend.

Research suggests certain vocabulary thresholds as determinants of learners' success or ability in using or understanding language. For example, it was reported by Nation (2001) that if individuals have receptive knowledge of the most frequent 2,000 word families, this enables them to understand 90 per cent of the vocabulary in conversations. According to Laufer (1992), knowing the 3,000 most frequent word families receptively is a prerequisite for the comprehension of authentic texts, while for Hirsh and Nation (1992), it is necessary to know more than 5,000 word families to enjoy reading. Therefore, some studies which have targeted the comparison of receptive and productive word knowledge of learners exist in the field (Fan, 2000; Laufer, 1998; Laufer & Paribakht, 1998). Yet, the test formats usually target receptive vocabulary size. The reason for L2 researchers to focus on the size or breadth aspect of vocabulary acquisition (how many words are known) might be that it is easier to measure and analyse this aspect than the vocabulary depth aspect (how well words are known), and a size test covering a large sum of test items is regarded more informative about students' overall state of vocabulary than a productive measure including limited words (Read, 2000).

As for L2 vocabulary acquisition, most models are based on the fundamental processes of child vocabulary learning described by Aitchison (2012) and follow the steps of labeling (attaching semantic meaning to the word form), packaging (fine-tuning the understanding of word meanings by establishing the extent of the association between the conceptual representation of the newly acquired lexical item and its phonological form as well as its further specification (e.g. while in the initial stages of vocabulary learning, any four-legged animal might be callled a dog, after gaining sufficient experience only certain animals are called dogs) and linking it to the lexical network, correspondingly.

In addition, L2 vocabulary acquisition shows some changes according to context of learning. In an EFL context, students make more effort to learn a similar amount of passive vocabulary than they do in an ESL context as vocabulary acquisition in the former one occurs mostly through form-focused instruction due to the paucity of language input while in the latter it is mainly based on more exposure. In other words, in either milieu, learners have advantages such as better activation of passive vocabulary in EFL contexts in accordance with the coverage of a course book used and larger passive vocabulary in ESL contexts as a result of exposure to a large number of lexical items, which also means there are different developmental patterns of vocabulary in different language learning contexts. In an EFL environment, while teachers sometimes pay attention to practising words in authentic and communicative tasks, some other times they may present them as decontextrualized items. It also seems effective to provide lower-level learners with L1 equivalents of L2 words to aid memorization. However, in general, presenting information to learners regarding word form, word meaning, word class as well as different word associations may help them create mental links and representations in their mental lexicon, which is defined by Jarema and Libben (2007) as “the cognitive system that constitutes the capacity for conscious and unconscious lexical activity” (p. 2). In this system, it is possible to store information and retrieve it, as it has an organized nature. Furthermore, the system is in a constant state of flux as its capacity allows for a variety of processes such as the possession, acquisition, conceptualization, use and loss of lexical knowledge during one’s life.

On the other hand, vocabulary acquisition is not only regarded important within teaching programs, it is also subject to testing. Langauge teachers should know that vocabulary tests can be used for different purposes: to check whether the words included in a course were acquired (i.e. achievement testing), to detect any gaps existing in individuals’ vocabulary knowledge (i.e. diagnostic testing), to put learners in suitable classes according to their level of knowledge (i.e. placement testing), or to make estimates of an individual’s skills representing his/her target language performance (i.e. proficiency testing). For example, in order to test whether learners’ are successful in acquiring the words covered in a particular course, it is expected that teachers select the test items from the course book. However, they should also know that there are different test options, such as vocabulary size tests, in which words to be tested are more general frequency-based samples and which they can use for both diagnostic and placement purposes. By adapting such tests, (e.g. using test items covered in course materials) it may be possible to use them as progress or proficiency tests during or at the end of a course since they have practical formats measuring different language abilities. In addition, teachers can compare the results of vocabulary size tests with the scores students get from examinations of course materials that are used in L2 classes as such a comparison can aid teachers while setting their future vocabulary size goals and deciding on the course materials as well as L2 teaching practices accordingly.

Today, there are several tests to assess receptive vocabulary size (e.g., Vocabulary Levels Test (Nation, 1983)). Such tests can be administered easily and marked quickly thanks to their numerical grading system, so they attract both teachers and researchers. In addition, all these measures have typically focused on the form-meaning link, at the recognition level in particular, as it is believed that if a word is recognized by learners, it means they know it. In addition, word recognition tests do not just tap into a small part of word knowledge. Vocabulary size estimates done through word recognition counts can in fact give information about the outer limits of vocabulary knowledge because of the fact that in order to understand or use words with any depth of meaning, first they have to be recognized (Cameron, 2002).

The purpose of the present study is to compare and contrast the scores obtained from three recognition-based receptive vocabulary size tests as there is a lack in the field regarding such studies. Most recent studies on vocabulary testing have focused on L2 vocabulary knowledge estimates through vocabulary size tests (e.g. Laufer 1998) as having sufficient vocabulary knowledge and being linguistically competent in a L2 appear to be strongly correlated. However, though vocabulary size tests have seen wide recognition and application in the field of vocabulary testing, they actually present testees with different tasks which measure different language abilities and level of vocabulary knowledge. As a result, it is expected from them to offer different size estimates, which means they may not be able to measure vocabulary knowledge to the same magnitude; an idea that is in strong opposition to prevailing opinions in the literature. In accordance with this idea, the first aim of the present study is to find out if the three tests used in this study in fact provide similar vocabulary size estimates as claimed in the literature. The second one is to examine which test correlates best with the exit score of the participants that reflects their academic success.

1.1. Background of the Study

The VST (Nation & Beglar, 2007a, 2007b), VLT (Nation, 2001) and Y/N Test (Meara, 1992) are often used in scientific studies which aim to test learners' receptive knowledge of vocabulary. The VST has the multiple choice format, VLT has the matching format, and the Y/N Test, which is regarded as a self-report, has the checklist format. It is assumed in literature that these three tests are equally powerful tools for working out an overall figure of receptive vocabulary size, and thus the results of different studies in which any one of these tests is used are considered to be comparable. For instance, a vocabulary size of the first 3,000 words measured by the Y/N test is accepted equal to a vocabulary size of the same magnitude measured by the VST. There are reasons, though, to argue they are not equal when their underlying constructs are taken into consideration. To put it differently, it does not seem possible to accept these tools as equal because while the Y/N Test targets meaning-recall (passive recall), the VST is a meaning-recognition (passive recognition) and the VLT is a form-recognition measure (active recognition), all being different strength modalities as offered by Laufer and Goldstein (2004). Though Laufer and Goldstein's study shows that passive recall is the best predictor of classroom language performance, it in fact seems possible to employ all these tests in a L2 classroom according to the activity used and knowledge type wanted to be activated or practised. They also appear to be helpful due to the insights they may offer for the cognitive processes involved in vocabulary acquisition and better understanding of the nature of vocabulary knowledge, especially in an EFL context.

A closer look at the abovemnetioned tests reveals that the Y/N test has the simplest format as lexical items are provided to examinees in isolation and they just indicate whether they have knowledge of each item or not (Schmitt, 2010). Therefore, it could be expected to produce higher size estimates than the others. In addition, because checklist tests require no direct knowledge demonstration, it seems very possible for examinees to overestimate their word knowledge as they cannot resist the tendency of checking items they do not actually know, so their precision in terms of how correctly they can determine vocabulary size or knowledge is open to question. In order to counterbalance any possibility of overestimation, non-words are included in Y/N tests, but according to Cameron (2002), it does not seem to be a good solution because as she argues, students might only partially understand words from first encounters, and they have to use this partial knowledge awaiting some further encounters to obtain supplementary information. Furthermore, non-words bear a resemblance to English words, so learners who have partially known words in their mental lexicons might have difficulty in distinguishing them from 'real' words, as similar sounds might become confusing within the phonological loop of the working memory (Baddeley, 1997). What is more, she states that the format of the Y/N test is simpler than that of the VLT and goes further stating that although the Y/N test like the other two tests uphold the idea of word frequency levels, except for the 2K as well as Academic lists, the levels used by Meara (1992) and Nation (1990) involve different lists, so they do not keep up an exact correspondence with each other. In her study, Cameron found that the VLT was more useful than the Y/N test for measuring the performances of secondary school learners, chiefly because of the inclusion of non-words as they produce unreliable results. After the frequency level based correlations of the two test scores were performed, only two significant correlation values were reported by Cameron: between Academic (Y/N) and 5K (VLT) words, and between 3K (Y/N) and 2K (VLT) words. This is a quite weak level of correspondence because two significant correlations between such tests might be found simply by chance. If these tests had had the same underlying construct, there would probably have been more instances of significant correlation, especially between the Academic and 2K words, since both measures tested the same words at this level. As a result, even though the Y/N test is preferred by many students as it is an easier measure to complete, it does not seem interchangeable with the VLT, which is somewhat more demanding. It is also argued that the Y/N is not a very functional test for low level or weak learners as certain correction formulae that are used to deal with testee's guessing behavior and response bias might reduce their scores because such people "respond unpredictably to the pseudowords" (Beeckmans, Eyckmans, Janssens, Dufranne, & Van de Velde, 2001, p. 240), and they usually misread items in unpredictable ways (Meara, 1996a). Another reported problem with non-words is underestimation, which means non-words might cause some testees to feel too conservative to check items which they know (see Eyckmans, 2004; Harrington & Carey, 2009; Mochida & Harrington, 2006; Stubbe, 2012).

On the other hand, according to some researchers, compared to the VLT, where answers might be guessed easily, the VST, with its multiple choice questions design, is said to ensure maximal reliability and validity (Nation & Beglar, 2007a; Wang & Du, 2014). However, regarding their formats, the VLT requires test-takers to recognise target word forms after they read the given short definitions, and the VST presents target items in short non-defining sentence contexts which are followed by four possible definition options and asks for meaning recognition. Therefore, though both tests provide evidence-based knowledge of each tested word (Read, 2007), the VLT requires a higher and stronger degree of knowledge of word meaning. This assumption is actually supported by some studies (Laufer & Golstein, 2004; Laufer, Elder, Hill, & Congdon, 2004). In fact, it is possible to use the abovementioned measures in language classes as they seem to have more practical designs than institutional exams prepared as formative or summative assessment tools. For instance, at the beginning of a language course, the Y/N test can be used as a placement test or at the end of a course, the more demanding VLT or VST can be used as achievement or proficiency tests as they seem more appropriate for tracing gaps in learners' vocabularies and providing feedback on their overall vocabulary gains.

When all these taken into account, it once again seems necessary to correlate the scores obtained from these tests with each other to see if they really offer similar estimates of vocabulary knowledge. In fact if they do, they can be replaced by one another in research studies or when used to meet different goals related to vocabulary testing. In addition, since these measures were occasionally subject to different studies correlating the vocabulary knowledge estimates with overal proficiency (e.g. Meara & Jones, 1988; Staehr, 2008), it would be better to correlate each one of the scores obtained from the three tests with the participants' preparatory class exit scores to see which test will offer the highest correlation value. If any one of the tests can offer significantly higher correlation with the participants' preparatory class exit scores, it can be suggested to the administration by the researcher to consider replacing the given test with the instutional proficiency exam prepared by the members of the testing committee.

1.2. Statement of the Problem

Nobody will oppose the fact that testing word knowledge is not a very simple issue as vocabulary knowledge is a complex construct because of its multidimensional nature (Henriksen 1999; Nation 2001; Read 2000). Therefore, as argued by Schmitt (2010), multiple measures of vocabulary are required in order to make correct estimates of the participants' vocabulary knowledge. However, it seems to be a mistake to consider all receptive recognition vocabulary knowledge tests equal because when the way students follow in each test (form↔ meaning) is examined in detail, it is obvious that the constructs they aim to measure are totally different. For example, if participants are required to pay attention to the word form and meaning respectively, it can be said that meaning recognition aspect is measured; however, taking the opposite direction means this time form recognition aspect is on target (Nation, 2001). Therefore, this study seems important to show whether vocabulary size tests used here can be considered as equal alternatives while testing learners' vocabulary knowledge.

On the other hand, it is known that though students might know the same words, their knowledge strength may vary to a great extent. Some can just say the word given to them exists in English, while some may also provide its definition as words that are completely unfamiliar sequences of letters/sounds at the beginning become functional units in the lexicon over time. In this case, their word knowledge cannot be accepted equal. This was the main point argued by Laufer as well as her colleagues (Laufer & Goldstein, 2004; Laufer et al., 2004). Also as language teachers, we usually favor practical multiple choice tests as they are easy to prepare and mark and students are familiar with the format because of our exam based education system. In addition, we generally tend to label students according the their exam grades and usually forget that the same testing tool may not be able to work equally well with all students as their word knowledge may vary to a great extent especially in an EFL context. Moreover, while preparing tests, we do not know how teachers decide on the assessment tool and whether they ask themselves why they choose a particular type of question. In fact, some simple changes we make in the question design can cause a big difference on the level of the cognitive ability required as the underlying test construct changes. Because of such issues, whether the tests used here should be accepted as equal measures seems rather controversial.

1.3. Purpose and Significance of the Study

There are ample studies in the area of vocabulary testing that provide convincing evidence of concurrent validity for the original forms of the redesigned tests used in this study; however, since researchers did not use identical items, content and format were usually confounded in these studies (e.g. Meara & Buxton, 1987; Shillaw, 1996, and most items in Cameron, 2002), and those whereby identical items were tested were limited to the comparison of performance on only two tests, mostly the Y/N test and VLT (see Eyckmans, 2004). In addition, it is obvious in the litearure that previous studies focused their analysis either on differences between the frequency level of the items (Mochida & Harrington, 2006) or on the proficiency levels of the universities (Stubbe, 2012), not on both as done in the current study.

The logic behind employing receptive vocabulary knowledge tests in this study is that both these tests and the vocabulary included in the course materials used in the institution where the study was carried out are based on frequency, and that when frequency-based vocabulary acquisition is the issue, both native and non-native language learners are commonly assumed to be following the order of range as well as frequency (Nation, 2006). Moreover, according to Meara (1996), when people deal with a small lexicon, which is composed of 5000 or 6000 words regarding English, vocabulary size dimension becomes really important. In fact, in an EFL context, we do not expect students to know a higher number of words than this sum. Furthermore, in order to fill the gap of the content validity mentioned above, the study uses multiple measures of the same test items through three adapted bilingual vocabulary tests with the same desing features as the VST (Nation & Beglar, 2007a/b), VLT (Nation, 2001) and Yes/No Test (Meara, 1992) and deals with the key question of whether the abovementioned three frequency-based vocabulary tests are equally successful at measuring the participants' receptive vocabulary size to the same extent and be accepted as equal alternatives while measuring word knowledge, which was not in the scope of any previous studies. Organised in this way, the study will probably offer different vocabulary size estimates and help paint a more complete picture of vocabulary knowledge as different test items are expected to tap different type of vocabulary knowledge in terms of receptive or productive mastery, and the results are expected to support the idea that each test measures a different word knowledge aspect at a different degree of mastery. In other words, although all the tests are usually labeled as passive recognition measures in the literature, the cognitive vocabulary knowledge process activated by test items, such as knowledge of meaning or form, is different, so the direct comparison of multiple assessment performance is expected to provide better construct-related validity evidence, which lacks in the field of vocabulary research. Above all, there are no previous studies which keep participants' prior vocabulary knowledge stable and purely focus on the comparison of the test formats, which will be done in this study. In fact, such a comparison is of vital importance because it is believed by the researcher that different test designs actually aim to measure a different aspect of word knowledge and a different strength modality. For instance, unlike the decontextualized Y/N format, the VLT and VST are cued recognition tasks that provide a partial context to students; therefore, in the current study, with their richer context which will encourage the use of partial knowledge, they are expected to result in better performance.

In addition, researchers like Nation (2006) argue that whereas the most frequent 2K words deserve classroom time to be learned, for lower-frequency words, it is better to teach students some vocabulary learning strategies and techniques to use outside classrooms to contribute to their vocabulary development. Also, since language proficiency and word frequency are regarded to be two important factors in vocabulary acquition, the study will focus on the question how they will affect the overall test scores.

Finally, since these measures are not considered the same in terms of their underlying constructs, the study will mainly focus on the correlation values they offer and the question which one will provide the highest correlation with the participants' university preparatory class exit scores. Correlation figures are important in order to be able to definitely say that tests can be used instead of each other not only for research purposes but also in language classrooms provided that they offer similar vocabulary knowledge estimates. Also, the test which offers the highest correlation with the students' university preparatory class exit scores can be taken as an alternative to the institutional proficiency exam.

1.4. Research Questions

Based on the key considerations mentioned above, the current study aims to address the following research questions:

1. Do the three English vocabulary tests, the Vocabulary Size Test (Nation & Beglar, 2007a, 2007b), Vocabulary Levels Test (Nation, 2001) and Yes/No Test (Meara, 1992), reveal similar overall receptive vocabulary knowledge estimates in different proficiency levels?

2. Do the VST, VLT and Yes/No Test reveal similar overall receptive vocabulary knowledge estimates across test sections?

3. Do the VST, VLT and Yes/No Test reveal similar overall receptive vocabulary knowledge estimates in different frequency bands for different proficiency levels groups?

4. a. How well do the VST, VLT and Yes/No Test correlate with one another?

b. Which of the VST, VLT and Yes/No Test has the best correlation with the participants' university preparatory class exit scores and yields consistent results?

1.5. Limitations of the Study

Although the study provides convincing evidence that researchers had better not accept the three size tests used here as equal measures, there are a number of limitations to the study suggesting cautious interpretation of its findings. First of all, the scope of the study was limited to solely one facet of vocabulary knowledge, namely vocabulary size, so simply the link between word form and word meaning was tested, though the importance of depth knowledge was acknowledged. Therefore, instead of multiple meanings of each word, only their single meanings could be measured.

In addition, beacuse of some administrative constraints, the study had to include only a limited number of students while seeking answers to the last research question, so it was not possible to correlate the exist scores of all the participants with their test scores.

Moreover, in one of the institutions, because of both institutional constraints and the participants’ personal preferences, it was not possible to include technology in the study. Therefore, pen and paper tests, rather than their online versions, were administered to the students there.

Another limitation of the study was that the study started with a higher number of participants; however, because of some technological problems, almost one third of them could not take part in the whole process.

On the other hand, the think-aloud protocol was carried out with a very limited number of students and only during the pilot study. It might have been better to add a qualitative aspect to this study by repeating the same protocol with more participants in the actual study.

This part discussed the importance of vocabulary knowledge, its relation to individuals’ academic success, vocabulary knowledge testing and the instruments used for this aim. Following the background of the study and the statement of the problem, the purpose and significance of the study as well as the research questions were presented. Finally, some information on the limitations of the current study concluded the chapter. The next part will review the related literature on various test constructs, three vocabulary tests which are compared and contrasted, the relation of frequency and language profiency to vocabulary knowledge, and lastly some sample studies including any of these tests. The third chapter will focus on the methodology of the study and deal with data collection tools and procedures and data analysis procedures. The fourth chapter will cover the results and the discussion of the findings in the light of the relevant literature. The last chapter will include overall conclusions, recommendations for pedagogical implications and personal suggestions for further research studies.

Chapter II

Review of Literature

In this section, firstly the term "construct" and three sets of major vocabulary knowledge constructs, namely size-depth, receptive-productive and recognition-recall, will be discussed. Next, Laufer and her colleagues' and Schmitt's classification of the form-meaning link will be addressed. These will help clarify and define the underlying constructs of the tests employed in the study. Then three vocabulary size tests used in this study will be described, and they will be compared and contrasted in detail. Following that, word frequency effect on vocabulary knowledge and vocabulary size effect on overall language proficiency will be dealt with. Lastly, studies which the current one is based on will be investigated, and this will be followed by some information on the purpose of this study and relevant research questions. All these will provide a theoretical basis for operationalizing the passive recognition dimension of word knowledge, claimed to be the same construct underlying the tests used.

2.1. Vocabulary Size Test Constructs

The term "construct" is predominantly a psychological term (Gyllstad, 2007); however, it has an extensive use in language testing as it forms the basis of both test choice and design and help clarify the actual meaning of scores obtained (Alderson, Clapham, & Wall, 1995; Bachman & Palmer, 1996; Chapelle, 1998; Ebel & Frisbie, 1991; Read, 2000; Read & Chapelle, 2001). According to Alderson (2001), in language testing, it is crucial to define our test constructs; that is, "what we are trying to measure" (p. 12), or in Bachman and Palmer's (1996) terms "what a given test task is supposed to be measuring" (p. 173) because as Tanaka (2012) says in the pursuit of a research project, without knowing your target construct, it is not possible to be on the right track.

As for the definition of "construct", according to Cronbach and Meehl (1955), it means any postulated attributes reflected in a test score (p. 178), and for Bachman and Palmer

(1996), it is a specific language ability which is based on an observable product and which consists of a score; it can even be a verbal description because the way we define construct will have consequences in terms of scoring. Regarding the fact that a learner might possess a specific attribute fully or to some degree at any time, or s/he might not possess it at all (Gyllstad, 2013), Chapelle (1998) outlines three approaches to construct definition one of which is the "trait definition" (p. 34) and supplies a complementary definition stating that "a construct is a meaningful interpretation of observed behavior" (p. 33). In vocabulary testing, a construct is defined by trait theorists mostly in terms of knowledge of vocabulary dimensions such as vocabulary size as well as fundamental processes like lexical access of a testee. A score on a particular vocabulary test, in this case, constitutes the observed behavior which is to be meaningfully interpreted because it indirectly shows some sort of mental ability or knowledge (e.g. knowledge about words). For Davies et al. (1999), constructs are particular traits that can be measured. More precisely, it is "an ability or set of abilities that will be reflected in test performance, and about which inferences can be made on the basis of test scores." (p. 31). Fulcher and Davidson (2007), on the other hand, define construct as "an attribute of the test taker" (p. 16), and for Fulcher (2010), it is the language ability underlying a test performance, though being far beyond direct observation (for similar definitions, see Eyckmans, 2004; Read, 2000).

One important point Alderson and Banerjee (2002) and Read and Chapelle (2001) emphasize about trait definition approach is according to trait theorists a test performance and an examinee's characteristics are interrelated; consequently, while designing a test to measure a particular aspect, they eliminate any kind of context as much as possible since contextual variables are regarded insignificant. The area of vocabulary testing is dominated by trait definitions and they are operationalized by means of assessment tools which are *discrete*, *selective* and *context-independent* (three dimensions proposed by Read, 2000, p. 9). In such tests, target items are presented in isolation. Discrete means vocabulary knowledge is measured as an independent or separate construct, selective means the test focuses on particular vocabulary items, such as words at specific word frequency levels, and context-independent means the expected response should be produced without any contextual data.

On the other hand, for a more precise construct definition, Bachman (1990, p. 40-45) offers a three-stage analysis. This means besides offering theoretical and operational definitions of a construct, it is necessary to establish procedures to quantify observations. When a construct is defined theoretically, it means the target ability and its characteristics are specified in a way that they do not bear any similarity to any other constructs, while operational definition attempts to make the construct observable. The last step, though, requires the quantification of observations using a particular scale. As Eyckmans (2004) exemplifies, while designing a vocabulary size test, the process of construct definition in terms of theory necessitates the clarificitaon of what receptive vocabulary means. In other words, it is necessary to determine what kind of language ability "receptive vocabulary knowledge" refers to. Then in the ESL/EFL context, a construct can be regarded as a language theory, whose operalisation is triggered by a test.

In accordance with the discussions so far, in this study, Chapelle's definition of construct, that is "a meaningful interpretation of observed [language] behaviour" (1998:33) will be followed. To be more specific, taking a trait perspective and regarding Schmitt's (2010) advice on specifying exactly what particular dimensions of vocabulary knowledge are addressed, the construct being measured in this study will be limited to the dimension *vocabulary size*. This dimension of vocabulary knowledge will be operationalised using three different frequency based receptive recognition tests (see Appendix 1) which will be discussed in detail in the remainder of this chapter, and response consistency of the testees across these measures will be looked for. However, before that, the distinction between three sets of major constructs of vocabulary knowledge will be presented.

2.1.1. Size (breath) dimension versus depth dimension of word knowledge. It is usually recognized that a word knowledge definition basically embraces the *breadth* and *depth* aspects (Anderson & Freebody, 1982), which have recently pre-occupied vocabulary testing (Akbarian, 2010; Chui, 2006). Many researchers state that "*breadth of vocabulary knowledge"* is usually represented as " vocabulary size" (Chen, 2011; Elmasry, 2012; Gyllstad, 2009; Henriksen, 2008; Read, 2004; Zareva, 2005). This aspect of word knowledge is defined by Nation (2001), Nergis (2013), Meara and Wolter (2004), and Read (2004, 2007) as the number or quantity of words known, and for Chapelle (1998: 36) it is "the absolute number of content words a person knows". According to some researchers, it might be some superficial knowledge or understanding (Marzban & Hadipour, 2012; Nadarajan, 2008; Qian, 2002; Qian & Schedl, 2004) that can be tested through a simple response task of indicating whether certain target words are known or not (Read, 2007). Therefore, while vocabulary size calculation sometimes hinges on the whole words recognised by individuals as existing lexical items, and they are not to attach any meaning or translation to them or use them with a great subtlety (Milton, 2010), some other times, it is expected from learners to understand the semantic features of the given word to a certain degree. For example, they must figure out that "bird" is a creature with wings and feathers and it can fly (Cooper, 1997: 110).

As for the dimension of "*depth of vocabulary knowledge",* which is beyond the scope of my thesis, some researchers consider it as the quality of knowledge or how good a lerner's knowledge is about a certain word (Anderson & Freebody, 1982; Chen, 2016; Choudhury, 2015; Meara, 1996a; Meara & Wolter, 2004; Nation, 2001; Nergis, 2013; Read, 2000; Schmitt, 2010; Wesche & Paribakht, 1996; Wolter, 2001). This dimension actually refers to the degree or strength of knowledge. For some, depth aspect has to do with knowledge of different aspects of a particular word (Bardakcı, 2016; Henriksen, 1999; Nation, 2001; Schmitt, 2014; Tanaka, 2012). Consequently, it is defined as the network building process through which learners create mental links among L2 words found in their lexicon such as paradigmatic relation of synonymy (Haastrup & Henriksen, 2000; Qian, 2002; Varnaseri & Farvardin, 2016).

On the other hand, although both dimensions have an equal importance while acquiring a language, there is inconsistency in the way the distinction between them is handled by scholars. For some researchers, this distinction is not so clear cut (e.g. Milton, 2009), and some others such as Zimmerman (2004), Vermeer (2001), and Li and Kirby (2014) assert that the division is not as distinct as it seems probably because they believe that size and depth are somewhat closely related concepts which can partially overlap and show parallel development. However, there are particular frameworks which the construct vocabulary knowledge is based on, and in which the distinction between depth and breadth knowledge is categorically reflected (e.g. Nation, 2001; Qian, 2002; Read, 2000; Wesche & Paribakht, 1996; Qian & Schedl, 2004; Hasan & Shabdin, 2016; Schmitt, 2014; Hatami & Tavakoli, 2012; Greidanus & Nienhuis, 2001; Wolter, 2001). This distinction, as stated by Meara and Wolter (2004), is also often made by researchers of vocabulary learning or testing (e.g. Read, 2000, 2004; Schmitt, 2014). For example, Laufer et al., who suggest that "we need separate estimates of both size and strength to fully understand the degree of a learner's vocabulary knowledge" (p. 224), advocates the necessity of such a distinction especially for diagnostic purposes. Emphasizing the importance of the recognition that size and depth aspects are interrelated, Dóczi and Kormos (2016) as well supports the idea that the distinction between them is useful in terms of pedagogical and research purposes, especially when the aim is assessing L2 vocabulary knowledge. Zimmerman (2004) also highlights that though recent studies have started to blur this distinction, there is still a tendency in literature towards the differentiation between both dimensions, especially when learners' general language proficiency is the issue.

Another important point is that the size dimension, usually defined as knowing the form or primary meaning of a word, is regarded as the most important vocabulary knowledge aspect. According to some researchers such as Laufer et al. (2004) and Laufer and Goldstein (2004), knowing a large sum of words along with their first meanings is considered more important than knowing a limited number of items with thier additional meanings and relations to other words. This logic highlights the rationale for why most test designs focus on vocabulary size aspect, rather than depth, as "they [size tests] can give a more representative picture of the overall state of the learners' vocabulary than an in-depth probe of a limited number of words" (Read, 2000, p. 115). Furthermore, regarding language proficiency, Ling (2015) notes that the breadth of vocabulary is much more indispensable compared to the depth of vocabulary, and Li and MacGregor (2010) state that "knowledge of a large number of words does not guarantee high language proficiency, but without vocabulary size reaching a minimum threshold, learners will be unable to successfully engage in either receptive or productive language use" (p. 239). Zhang (2013) shares the same belief noting that if a student knows a satisfactory sum of words, this allows him or her to comprehend written materials or conversations in an unaided way. Henriksen (2006) and Schmitt (2010) also posit that the role of breadth knowledge is significant in L2 proficiency in general, and Zhang, Pan, and Xu (2014) emphasize how great the impact of vocabulary size on English abilities is. Meara (1996a) as well stresses that size is the major aspect of L2 lexical knowledge, and according to him, under equal conditions, the students possessing bigger vocabularies are accepted more proficient than the ones with smaller vocabularies, as supported by evidence suggesting that vocabulary skills contribute to almost all L2 proficiency aspects.

Taking all these into consideration, it is clear that one of the vocabulary constructs that have attracted a lot of attention in vocabulary testing is vocabulary size. Similarly, the focus of the current study is on the same construct, and following Read's (2000) definition, it refers to "the number of words that a person knows" (p. 31).

2.1.2. Receptive versus productive knowledge. The long-standing division between receptive and productive dimensions, which is considered to have connections with the division between size and depth aspects (e.g. Cooper, 1997), has been made at least a century ago (Milton, 2013), and is claimed to be the best known dimension to "knowing a word" (Laufer & Nation, 1999). That is why it takes place in some taxonomies which demonstrate many dimensions and degrees of knowledge. For example, according to the taxonomy proposed by Henriksen (1999) and Haastrup and Henriksen (2000), vocabulary knowledge has three dimensions: 1) from partial towards precise, which reflects the meaning comprehension improves; 2) depth of vocabulary knowledge, which shows similarity to network building with representations of word associations; and 3) from receptive to productive knowledge, which shows the learner's control of word knowledge and access to it. Alternatively, Nation (2001:27) offers a definitive list encapsulating what knowing a word means. He perceives vocabulary knowledge as a construct comprising a variety of aspects. The first one is "knowledge of form". Next comes "knowledge of meaning" aspect, which is followed by the final aspect of "knowledge of use". In addition, each aspect is composed of some further dimensions of knowledge and can be acquired not only receptively but also productively (for similar taxonomies, see Chapelle, 1998; Cronbach, 1942; Meara, 1996a; Nation, 2001; Ozturk, 2003; Qian, 2002; Richards, 1976). This common receptive-productive division is well recognised in L2 teaching by a great number of researchers and has been subject to many studies (e.g. Bardakcı, 2016; Coxhead, Nation, & Sim, 2015; Dewaele, 2004; Gyllstad, 2004; Mondria & Wiersma, 2004; Read & Chapelle, 2001; Read & Nation, 1986;

Stewart, 2012; Waring, 1998; Zareva, Schwanenflugel, & Nikolova, 2005; Zimmerman, 2004).

In order to emphasize the ecological validity of receptive-productive dichotomy, Schmitt (2010) mentions that most language teachers experience that they have students in their classrooms who understand lexical items pretty well while listening to a recording or reading a text, though they are not capable of producing them in a spoken or written context. Additionally, Read (2000) and Azodi, Karimi, and Vaezi (2014) state that as language learners, we can realize that the number of words we can recognize and comprehend when we read or listen to them is relatively larger than the number of those we are able to use during speaking or writing activities. The common characteristics here is that the former construct calls for word form recognition and retrieval of its meaning through a receptive skill, while the latter is the ability not only to recall but also to produce the proper word through a productive skill.

In literature, the receptive-productive dimensions are usually equated with passive-active dimensions (Nizonkiza & Van den Berg, 2014), and the terms *receptive*/*productive* and *passive*/*active* are usually used by some researchers interchangeably as it will be done in this study (e.g. Corson, 1995; Laufer, 1998; Milton, 2009; Read, 2000; Schmitt, 2014). These two dimensions of vocabulary are operationally defined in numerous ways by researchers. For example, receptive knowledge is regarded by Waring (1997), Takala (1985), Laufer and Goldstein (2004) and Webb (2008, 2009) as being able to translate the tested L2 item into the first language word, which means a learner can perceive the target item and retrieve the meaning of it (L2 to L1), and productive knowledge as being able to follow the prompt given in the first language and produce its equivalent in the target language, which shows the learner is able to retrieve the target L2 word that expresses a specific meaning and has control over its spelling (L1to L2). Laufer et al. (2004) further develop the concept suggesting that receptive knowledge means “supplying the form for a given concept”, whereas productive knowledge is “supplying the meaning for a given form” (p. 206). Such definitions have a connection with the development and desing of different vocabulary tests. For example, while testing passive word knowledge, sometimes researchers ask for the L1 equivalent of the target word, some other times they ask the testee to mark the correct option from among the given meaning or form alternatives accordingly for the tested word form or meaning (see Waring, 1997). In order to test active knowledge, though, they ask for the L2 equivalent of the L1 test item (see Takala, 1985), or alternatively they ask for the correct use of a target word in an original sentence. In such definitions, it is obvious that receptive-productive dimensions are restricted to meaning-form aspects, which will be dealt with in the following sections. It also supports the findings of some researches showing that different vocabulary-item types are able to tap different aspects and different degrees of vocabulary knowledge (e.g. Laufer & Goldstein, 2004; Laufer et al., 2004; Ozturk, 2007; Schmitt, 2010; Webb, 2008, 2009).

Another issue which causes division between scholars regarding the notion of the abovementioned distinction is the fact that there is “no consensus as to whether this distinction is dichotomous or whether it constitutes a continuum” (Laufer & Goldstein 2004: 405), or in Shah, Gill, Mahmood and Bilal’s (2013: 42) terms, is this distinction “bipolar”or “binary”? For some scholars, like Melka-Teichroew (1982) it is bipolar. Melka assumes that when a word is learnt, the two dimensions can be regarded as two levels on the same continuum, which means lexical knowledge has a kind of developmental pattern. In other words, the distance between these dimensions is thought as increased familiarity of a particular word; therefore it has widely been accepted by many researchers that information load of a learner accelerates the gradual shift from receptive toward productive mastery (e.g. Alkhofi, 2015; Eyckmans, 2004; Hayashi & Murphy, 2011; Henriksen, 1999; Read, 2000).

The theory stems from the idea that word knowledge cannot be considered as an "all-or-nothing" phenomenon. On the contrary, it ranges on a continuum including different aspects and degrees of strength (Laufer, 1998: 256; Laufer, et. al., 2004: 209; Schmitt, 2000: 6). Read (2000) and Ozturk (2016) explain this as follows: When students come across a new word, they gain so limited knowledge that they might not be able to recall it until a second encounter. Once they have gained more knowledge of the word's spelling, pronunciation, meaning, grammar, range of its use, and so on, thanks to repeated encounters, then it will be possible for them to use that word themselves. Simply put, productive knowledge of vocabulary builds on receptive knowledge (Zhou, 2010), since research reveals that vocabulary learning proceeds along a continuum and knowledge of vocabulary develops better in a longitudinal setting (Haastrup & Henriksen, 2000; Ozturk, 2015). The developmental continuum of receptive and productive mastery also backs up the notion by Henriksen (1999) and Haastrup and Henriksen (2000) that word knowledge follows a specified direction; heading from zero toward partial and precise. That is, it moves "from recognition to vague understanding of the meaning and later to the mastery of a precise comprehension" (Zhong, 2011: p, 118), or as Schmitt (2000) states, it progresses from unknown to knowing and eventually to full mastery. Also, development in one aspect, it will inevitably trigger development in the other two. The gradual movement, according to Laufer and Goldstein (2004) though, starts with passive recognition, and goes on with the modalities of active recognition, passive recall and active recall, respectively, which will be given further attention in the next section.

In opposition to Melka, Meara (1990; 1996b; 1997) states that since these two vocabulary knowledge types represent different associational knowledge, they cannot form a continuum. For him, degree of automaticity is the determinant of either knowledge domain. According to the lexical organisation proposed by Meara (1990), words that are known productively are connected with a productive item in the lexicon, which makes it accessible for learners, while the ones that are known receptively lack an incoming link from the lexicon. That is why they cannot recall them unless an external stimulus like hearing or reading activates them. Then, items which have the right type of connection in a learner's mental lexicon would eventually become productive, (that is, active vocabulary may exist on a continuum), whereas the ones that lack such connections would unfortunately remain at the receptive level (that is, passive vocabulary may not exist on a continuum). Based on such opinions, it can be said that this dichotomy-like phenomena does not have a clear-cut nature. Relatedly, Schmitt (2010) posits that we cannot say for sure whether these knowledge types form a continuum or not because though lexical items are typically assumed to be known either in a receptive or both receptive and productive manner, in reality, for any word, a different type of knowledge can be gained at varying receptive and productive degrees. Here, it is significant to point out that though the abovementioned distinction is considered to be one of degree of knowledge rather than absolute, in accordance with the aim of the present study, it will be reflected in dichotomous terms.

On the other hand, despite the problems with the conceptualization and definition of receptive and productive knowledge, these two constructs are typically in the scope of vocabulary testing, and the following conclusions about them are drawn in literature: a) in terms of size, receptive vocabularies are typically larger than productive vocabularies (Choudhury, 2015; Fan, 2000; Henriksen, 2013; Koya, 2005; Laufer, 1998, 2013; Laufer & Paribakht,1998; Laufer & Waldman, 2011; Meara & Fitzpatrick, 2000; Nemati, 2010; Sakai, 2009; Schmitt & Meara, 1997; Shin, Chon, & Kim, 2011; Tschirner, 2004; Waring, 1997; Webb, 2008; Zhong & Hirsh, 2009; Zhou, 2010; Zimmerman, 2004); b) compared to productive vocabulary, receptive vocabulary grows faster, with the gap between the two getting wider as learning progresses (Laufer, 1998; Laufer & Paribakht, 1998; Webb, 2008);

c) receptive knowledge typically precedes productive knowledge (Ellis & Beaton, 1993; Fan, 2000; Horst & Collins, 2006; Laufer, 1998; Laufer et. al.,2004; Laufer & Paribakht, 1998; Lee & Muncie, 2006; Meara & Fitzpatrick, 2000; Milton, 2009; Mutlu & Kaşlıoğlu, 2016; Swain, 2005; Waring, 1997; Webb, 2008).

When it comes to mapping out a conceptual framework for this research, the subjects are not to produce anything, so the focus of the investigation will be measuring their receptive knowledge using different vocabulary size tests. Therefore, adopting Nation's (2001, p. 359) definition for receptive vocabulary items as the ones that "involve going from the form of a word to its meaning" and productive vocabulary items as those that "involve going from the meaning to the word form", I define receptive knowledge here as being able to understand the target word form acting as a stimulus in the stem for the given meaning options and productive knowledge as being able to understand the given meaning first and then hunt for the matching target word form.

2.1.3. Recognition versus recall. Some researchers, such as Nation (2000, 2001), state that during language learning process another set of key factors that affect learning difficulty is recognition and recall, which assumed to be distinct knowledge types. In the literature, this two-fold distinction is advocated by many by L2 vocabulary researchers and test designers (see Brown, Waring & Donkaewbua, 2008; Greidanus et.al., 2004; Griffin, 1992; Gyllstad, Vilkaite, & Schmitt, 2015; Jordan, 2013; Nemati, 2010; Schmitt, 1998, 2010; Tschirner, 2004; Vermeer, 2001; Waring & Takaki, 2003; Zhang, 2013; Zhong, 2018).

While the recognition-recall division is frequently substituted with the receptive-productive distinction (Eyckmans, 2004), Read (2000, p. 155-156) prefers to define these constructs in a more narrow scope using the terms "recognition" and "recall" as well as "comprehension" and "use" (p. 154-157). Recognition is triggered in tasks where learners are given an L2 word item and expected to demonstrate that they know its meaning by translating the word into L1, whereas in recall tasks, they are to recall and provide the target L2 item from memory with the help of a kind of stimulus presented to them in the form of L1 translation. On the other hand, Read (2000) defines comprehension as being capable of understanding a word in receptive contexts, while for him use refers to the ability of using the target word in productive tasks, such as oral retellings, translations and picture description activities. As understood from the definitions, for Read, the term recognition refers to the recognition of meaning and the term recall is used by him as the referent of form recall only. Seeing these definitions limited, Nation (2001, p. 359) offers the following definitions: "a recognition vocabulary item format involves the use of choices", whereas "a recall item requires the test-taker to provide the required form or meaning". The difference between the definitions offered by Nation and Read is related to the type of knowledge a test item requires. For example, translations from L2 into L1 are regarded as recognition based items by Read. Nation, however, states that a task of translation in either way aims to measure recall knowledge as a test taker is required to recall either an L2 word meaning and provide its L1 form or to recall and provide an L2 item form that matches a given L1 meaning. Likewise, when learners are provided with the L2 word form (as in L2→L1 translations), Read restricts the term recognition to meaning recognition only, while Nation suggests this term for both meaning and form recognition. In this case, as stated by Ozturk (2007), it is very obvious that Read's division of these two terms disregards certain multiple-choice format tasks. However, Nation's division applies to all multiple-choice format tasks and labels them as recognition measures, as agreed by some other researhers like Mochizuki (2012). Therefore, in agreement with Nation, a recognition test item in this study will refer to the one that requires recognition in both ways. In addition, I view form and meaning recognition as shared constructs of multiple-choice or matching formats which are not of the same difficulty since the items of form recognition and those of meaning recognition suggest a difference in the type of knowledge being measured. For that reason, I will use the terms receptive recognition and productive recognition offered by Nation and define the former as the skill of understanding first the given word form within the stem and then recognising its meaning provided in the options (word form→word meaning), and the latter as the skill of understanding first the given word meanings and later recognising the corresponding word form of each from among the options (word meaning→word form).

2.1.4. The form-meaning link. According to the traditional definition of word knowledge, a word item is regarded as known when its meaning is known (Makarchuk, 2013; Webb, 2008). Therefore, the basic form-meaning link has widely been accepted as the central component of vocabulary knowledge (Laufer et al., 2004; Laufer & Goldstein, 2004; Nation, 2001), and in vocabulary size measures, words are accepted as "known" when meanings and forms are associated with each other in a correct way (Levitzky-Aviad & Laufer, 2013). However, it seems illogical for some researchers such as Laufer and her colleagues to accept word knowledge as an all or nothing phenomenon. According to them, though learners may have knowledge of a certain word meaning, the level of their knowledge might vary, so they offer the dichotomous active-passive as well as recall-recognition distinctions and accordingly differentiate between degrees of strength (Laufer & Goldstein, 2004; Laufer et al., 2004). Laufer and Goldstein define these meaning-form association based parameters as follows: "supplying the form for a given meaning versus supplying the meaning for a given form and being able to recall versus only being able to recognise (whether form or meaning)" (2004: 405-406).

The first division suggests that knowledge level of individuals is not the same. For example, some people can remember an L2 item to express a particular meaning ("active" knowledge), while some others might lack this ability but can recognise the meaning of an L2 item when it is presented to them ("passive" knowledge). The second distinction tells apart individuals in a similar way. According to it, some individuals can remember the form of the tested item or its meaning, while some others might be unable to do this, but has the ability to recognise the target L2 form or meaning from among several alternatives. These divisions bring about four different modalities of strength of word knowledge, shown in Figure 1: a) “active recall”; b) “passive recall”; c) “active recognition”; d) “passive recognition”.

	Recall	Recognition
Active (retrieval of form)	1. Supply the L2 word	3. Select the L2 word
Passive (retrieval of meaning)	2. Supply the L1 word	4. Select the L1 word

Figure 1 Degrees of Vocabulary Knowledge (Laufer and Goldstein, 2004:407)

In “active recall”, if the test is a monolingual one, participants are required to supply correct L2 word form to show that they have understood the given L2 meaning. In a bilingual version, the L1 translation equivalent of the tested item is provided to test takers as the prompt, and they are required to supply its L2 form. Sometimes clues are given to testees, such as providing the first letter of the target word, in order to help them eliminate non-target items that have the same meaning; In “passive recall” monolingual tests, it is required from the examinees to show their understanding of the target word meaning by making them complete a phrase or short sentence which includes the tested item. In a bilingual version, participants are to demonstrate their knowledge of the L2 word meaning. In such tests, the prompt provides the target form in L2, and they are to translate in into L1 by paying attention to the given first letter. When it comes to the task in “active recognition”, in a monolingual test, the prompt which defines the tested item is in the target language, while in a bilingual one, participants are given the prompt as the first language translation of the target word. In both versions of these multiple-choice format tests, they must recognize and select the target L2 word form from among L2 distractors, which lack a semantic relationship but which are equally difficult because of belonging to the same frequency band. Although in such a recognition task production is not required either in speech or writing, Laufer and Goldstein (2004) consider it active based on the definition of *productive* or *active knowledge;* that is to say, “knowledge that is used in speaking and writing, and involves going from the meaning to the word form” as offered by Nation (2000, p. 446 / 2001, p. 359). According to Nation, it is possible to check active-productive knowledge through a recall task in which learners have to produce the target word, as in translating an L1 word into L2, or through a recognition task in which they have to recognize and select the target word from among four options; Lastly, in “passive recognition” tests, both in a monolingual and bilingual one, the target word (L2 form) is provided as a prompt. Its meaning is chosen from among four options which are either definitions (as in the monolingual version) or L1 translations (as in the bilingual one) of the distractors used in the active recognition mode.

Given below in Figure 2 are two examples illustrating the test used by Laufer and Goldstein (2004) and showing how each knowledge type can be elicited when testing the same target item “*melt*”. The monolingual examples are from Laufer et al. (2004) and the bilingual ones are from Nation and Chung (2009).

	monolingual	bilingual
active recall	Turn into water *m* _____	*m*_____ mencairkan
passive recall	When something *melts*, it turns into _____.	Translate the following words into Indonesian. <u>*melt*</u>
active recognition	*Turn into water* a. elect c. melt b. blame d. threaten	Select the correct translation for the following words. *mencairkan* a. elect c. melt b. blame d. threaten
passive recognition	*Melt* a. choose c. make threats b. accuse d. turn into water	*Melt* a. menolong c. memeriksa b. mencairkan d. memandang

Figure 2 Monolingual and Bilingual Item Types Testing Degrees of Vocabulary Knowledge

The key point here is that while test formats in these two levels are similar (i.e. multiple choice), the level, or strength, of word knowledge is supposed to constitute a difficulty based hierarchy. For example, Laufer and Goldstein’s (2004) study confirms that vocabulary skills are hierarchic as the modalities used are implicationally scaled. Active recall is the hardest one as it requires the production of word form for the given meaning, and it is followed by passive recall which is not as difficult because this time word form is given and what is required from the learner is to produce its meaning. On the other hand, the next two modalities are recognition based, so they are simpler. Active recognition appears third, and the last modality is passive recognition, ranking as the easiest. Laufer et al. (2004) agree with Laufer and Goldstein on the existence of a difficulty hierarchy; however, they show in their monolingual test that there is no significant difference between recognition-based modalities. Therefore, in some studies, except for the "passive recognition", the remaining three modalities are used (e.g. Laufer & McLean, 2016; Sonbul & Schmitt, 2009), and the difficulty hierarchy is supported. Stubbe (2014), testing passive recognition and passive recall, and Webb (2008), testing form recall and meaning recall, are also the advocates of the aforementioned hierarchy. Consequently, since word knowledge develops cumulatively and at a different pace, Laufer and her colleagues highlight the great value of having tests which show the hierarchic change in strength of word knowledge (Laufer & Goldstein, 2004; Laufer et al., 2004). The underlying message here is again that "there is a variety of test formats that could be used and which differ from each other in difficulty" (Nation & Chung, 2009: p. 556).

On the other hand, though the abovementioned categorization seems important and advantageous in terms of helping label the tested aspect as the one that measures receptive or productive knowledge, according to Schmitt (2010), the terminology tends to be confusing because rather than addressing the distinction between active-passive mastery, the categories focus on the question which word-knowledge elements participants are supplied with and which are elicited from them. Therefore, he tries to make the distinctions far more useful by describing them in much more transparent terms explained below.

2.1.4.1. Schmidt's categorization of the stages of form-meaning link. Giving the focus on form and meaning, Schmitt (2010) covers Laufer and Goldstein's (2004) categories and relabels them using the terms "form recall, form recognition, meaning recall, and meaning recognition" as illustrated in Figure 3. According to Schmitt, these labels make the construct that is measured more obvious, not only in terms of "what aspect" it requires but also in terms of its "degree of mastery" (p. 86) such as recall versus recognition.

Word knowledge	Word-knowledge tested	
Given	Recall	Recognition
Meaning	Form recall (supply the L2 item)	Form recognition (select the L2 item)
Form	Meaning recall (supply definition/L1 translation, etc.)	Meaning recognition (select definition/L1 translation, etc.)

Figure 3 Levels of Mastery of the Form-Meaning Link (Schmitt, 2010:86)

The following subsections exemplify the form of testing the target item "dog" (taken from Schmitt, 2010) according to these four degrees of word knowledge through bilingual tests (L1=German [*hund*]; L2=English [*dog*]): In "form recall", the meaning is given as an L1 equivalent, and the target word form in L2 is required ("Active Recall" in Laufer & Goldstein). (d __ *hund)*; In "form recognition" the meaning of a word is provided in L1, and learners are expected to recognise and select its L2 form from the given set of options ("Active Recognition" in Laufer & Goldstein). (*hund* a. cat b. dog c. mouse d. bird); In "meaning recall", the learners see the target word in L2 form, and they must supply its L1 meaning ("Passive Recall" in Laufer & Goldstein). (dog *h* __); In "meaning recognition" after the learners see the target word in L2 form, they are supposed to recognise and select its L1 meaning from several options ("Passive Recognition" in Laufer & Goldstein). (dog a. *katze* b. *hund* c. *maus* d. *vogel*)

In relation to the discussion so far, in this study, "receptive" or "passive" mastery of form and meaning aspects of vocabulary knowledge, at recognition level, is measured. Having decided to test the participants' vocabulary knowledge through the basic form-meaning link and regarding Nation's (2000) advice that "in experimental research, it is very useful to test the same word in several different ways" (p. 581), three most common size tests are employed in this study. Being recognition tests, they are used for measuring receptive aspect, like having the ability to recognise a word with reference to its primary meaning, "e.g. "solution" as in "solution of a problem" instead of "chemical solution" (Laufer, 1998:257). By comparing and contrasting the scores of the examinees through these three tests, the aim is to find out whether or not they measure the same construct, which might offer a good compromise for future research ventures within the field of vocabulary. The tests along with some information on their history of development and design features will take part in the next part following a chronological order.

2.2. Vocabulary Size Tests

2.2.1. The vocabulary levels test. This measure, also known as the Levels Test, was called by Meara (1996a: 3; 1994: 9) "the nearest thing we have to a standard test" and is believed by researchers to hold this distinction even today (e.g. Schmitt, 2010) although having gone through some iterations (Beglar & Hunt, 1999; Laufer & Nation, 1999; Schmitt et al., 2001). Though the primary aim of the scholar in designing this diagnostic tool was to meet pedagogical needs, the test is widely used even today by researchers to assess receptive vocabularies in ESL/EFL contexts (see Batista, 2014; Cobb, 1997; Elmasry, 2012; Li, & MacGregor, 2010; Nassaji, 2006; Schmitt & Meara, 1997; Shen, 2008; Shiotsu & Weir, 2007; Stæhr, 2008; Zhang, 2013).

2.2.1.1. Historical background of the VLT. This tool was formed by Nation in 1983. Since then, the test has undergone through some modifications and validation. In his initial validation work, Read (1988) found that the test was reliable and it offered an implicational pattern, which means if a learner shows the expected performance at a lower-frequency band, s/he could be assumed to have mastered higher frequency vocabulary as well, since vocabulary acquisition is stated to have a strong bond with the frequency of order. In 1993, Schmitt added three more versions to his original test, but they were not validated. Since more moves were necessary to explore how reliable and valid the test was, combining specific parts of the four versions, Beglar and Hunt (1999) created a new measure and validated this reduced test by administering it to 496 Japanese students. Consequently, it was proved that the VLT was reliable and valid; it was basically measuring a single construct; and the students' scores on the levels tested correlated with their TOEFL scores, supporting the concurrent validity of the test. In 2001, Schmitt et al. carried out another comprehensive validation study, where the four forms of the test were combined into two versions and multiple forms of validity evidence was used. The validation analyses proved that both versions were valid, and though not being equivalents, they produced similar scores. Version A of these more thoroughly researched forms of the VLT is available in Schmitt (2000, p. 192–200) and Version B can be found in Nation (2000, p. 676–695; 2001, p. 416–424 (the updated form of Schmitt et al.'s (2001) test) and Schmitt et al. (2001, p. 82–88). Both versions, which are treated by Nation (2001) as equivalent forms, are also found in Schmitt (2010). Besides these revised and expanded forms of the VLT, a productive version of it was created by Laufer and Nation (1999). In later attempts, it became possible for Laufer and Goldstein (2004) and Laufer et al. (2004) to develop a computerized version to test breadth and depth aspects of knowledge simultaneously. In 2007, Xing and Fulcher took a step to check for how reliable the A & B versions of the VLT were and found that the latter version was more reliable at the 5000 word level. In the following years, Mclean, Kramer, and Beglar (2015a) created and validated the Listening Vocabulary Levels Test (LVLT), which had no difference from the VST in terms of its desing features. In the same year, Mclean and Kramer (2015) created the New Vocabulary Levels Test (NVLT), which was the written receptive format of the LVLT (see www.lvlt.info), and a year later, they wrote a Japanese variant of the NVLT (Mclean & Kramer, 2016). Today, the VLT, which has been rather influential in its original as well as modified versions, is a common test in vocabulary research. It is also used for measuring general/academic vocabulary size of L2 learners and placement purposes.

2.2.1.2. The design features of the vocabulary levels test. The test appeared in the field as a diagnostic instrument: to identify the kind of vocabulary that a teacher should focus on and help students with vocabulary learning by determining the frequency levels they need to study (Nation, 1990, 2000, 2001; Read, 2000; Schmitt et.al., 2001). The test is composed of various sections based on frequency and measures students' knowledge of words representing different frequency levels. Therefore, it provides frequency profile information of language learners' vocabularies instead of functioning as a single-figure measure of size (see Cameron, 2002; Gyllstad, 2007; Schmitt et al., 2001). As a result, researchers such as Schmitt (2010) and Gyllstad et al., (2015) argue that it does not strictly measure vocabulary size; however, in studies, as a common approach, the frequency levels are combined to figure out a total size figure (for examples, see Alkhofi, 2015; Culligan, 2015; Laufer & Nation, 1999; Lemmouh, 2010; Liu, 2016; Mclean & Kramer, 2015; Schmitt & Meara, 1997; Van Hout & Vermeer, 2007).

This test aims to assess knowledge of word families at distinct frequency bands (2000, 3000, 5000 and 10000) and also addresses a section to test knowledge of academic vocabulary, whose items are not frequency-based and whose interpretation should not be made together with other levels. In the 1990 version, target words for this section were taken from the University Word List (UWL; Xue & Nation, 1984), while in the more recent 2001 version, they were sampled from the Academic Word List (AWL; Coxhead, 2000).

In both the original VLT and its parallel versions (A–D), there are six clusters per level. Each one consists of six target words given as options on the left and definitions or synonyms for one half of them are given as stems on the right, so with a total of 90 test words, the VLT takes about 20 minutes to complete (see Batista, 2014). Participants must make correct matches between the definitions and corresponding target words by writing or typing the number of the target word next to its meaning. Below is provided a cluster from 2000 word level as an example (see Figure 4):

1 arrange
2 develop _______ grow
3 lean _______ put in order
4 owe _______ like more than something else
5 prefer
6 seize

Figure 4 A sample cluster from the VLT (version B, Nation 2001: 416-424).

In the revised versions (e.g. Nation, 2000, 2001), though, with 156 target items, the test has 10 clusters measuring four 1000-word (K) frequency bands (2,3,5,10) and 12 clusters to measure academic vocabulary from the AWL. Similarly, each of the updated versions by Schmitt et al. (2001) consists of 10 clusters per section and 150 items in total. With additional clusters, the bigger and updated versions of the VLT take about 30 minutes to complete (see Laufer, 1998; Li & MacGregor, 2010; Hirsh, 2015; Schmitt et al., 2001). These tests are available for free on the personal websites of Paul Nation (printable, http://www.victoria.ac.nz/lals/about/staff/paul-nation), Tom Cobb (printable and online, http://www.lextutor.ca/tests/) and Norbert Schmitt (printable, www.norbertschmitt.co.uk).

Regarding the desing features of this test, some researchers judge the VLT to be in a matching format (e.g. Akbarian, 2010; Batista, 2014; Elmasry, 2012; Gyllstad, 2007; Kremmel & Schmitt, 2018; Milton, 2009; Pignot-Shahov, 2012; Schmitt, 2010; Schmitt et al., 2001; Tran, 2009; Webb & Sasao, 2013). However, for some others, it is a multiple-choice tool where each stem has six possible options (e.g. Cameron, 2002; Gyllstad et al., 2015; Mochida & Harrington, 2006; Stewart & White, 2011; Varnaseri & Farvardin, 2016).

The VLT measures receptive knowledge of vocabulary at recognition level. It aims to test how able a testee is to comprehend the meaning definition of an L2 item and recognise the matching form. Though some researchers argue that the test measures the very first level of the form-meaning connection (e.g. Amirian, Salari, Heshmatifar, & Rahimi, 2015; Kremmel & Schmitt, 2018; Schmitt, 2010; Schmitt et al., 2001), according to some, such as Milton (2009), the format of the VLT calls for a type of knowledge that is beyond passive recognition. In fact, the VLT is defined by Schmitt (2010) as “a form recognition test” (p. 197), measuring the degree of knowledge labeled by Laufer and her colleagues as active recognition.

Furthermore, the VLT is a decontextualized test as it presents the tested items in isolation, and similarly, definitions do not include any contextual clues. Although pedagogically context is important, such practical diagnostic measures can be used in ELT classes to identify the kind of vocabulary and frequency level to focus on and help students with the words they need to learn through some vocabulary activities. In the VLT, to help ensure comprehensibility, the short definitions written with simple vocabulary do not require any syntactic or grammatical sophistication or higher reading skills or comprehension (Read, 2000), which also means the only obvious linguistic feature addressed is vocabulary knowledge (Read & Chapelle, 2001; Schmitt et al., 2001). In addition, since distractors do not bear any relation among themselves either in form or meaning, even students with a little knowledge can make the correct response. It, therefore, provides data on whether testees know the first and most frequent meaning of the tested item as mentioned before. Moreover, apparently, each cluster aims to test three items, but as stated by several people, examinees also need to know the meaning of the distractor words whose frequency band is the same as the tested items in order to discard them (Mclean & Kramer, 2015; Nation & Beglar, 2007a; Schmitt, 2010; Schmitt et al., 2001; Stewart & White 2011). This means testees actually deal with six words in each cluster (Gyllstad et al., 2015; Schmitt, 2010).

In the initial VLT (Beglar & Hunt, 1999), each frequency band had a fixed ratio reflecting the word class distribution in English: 5 noun clusters, 3 verb clusters and 2 adjective clusters; however, in the revised versions (Schmitt et al., 2001), there were 3 noun, 2 verb and 1 adjective clusters. In addition, there is not a mix of word classes in the clusters.

In addition, the test offers some advantages in terms of practicality, administration and scoring. While calculating a learner’s vocabulary size, it is also possible to follow different procedures (for examples, see Cameron, 2002; Laufer, 1992, 1998; Read, 1988; Schmitt & Meara, 1997). In the current study, the total vocabulary size approach will be employed and the calculation of the scores will be done according to the method suggested by Schmitt and Meara (1997) and Read (2000), which is adding up the totals of each frequency level.

2.2.2. The yes/no vocabulary test. This test wich was developed by Meara and Buxton (1987) is another widely applied measure of vocabulary knowledge. It is regarded by some people as a very authoritative tool (Verspoor & Cremer, 2008) and preferred by some because of being so applicable for assessment and research purposes (Eyckmans, 2000). While acknowledging the role construct irrelevant factors can play, it is also stated to be a valid and potentially functional test (Read, 2000).

2.2.2.1. The historical development of the yes-no test. The origin of this test format dates back to a very simple format known as the “checklist”, where students are given a group of target items out of context and the only thing they have to do is to mark the ones they know (Beeckmans et al., 2001; Read, 2007). However, compared to the VLT and VST, which prove that tested items are known by examinees, such a self-report does not verify whether or not learners have actual knowledge of the target word meaning (Batista, 2014; Eyckmans, de Velde, van Hout, & Boers, 2007; Read, 2000). Such a self-assessment instrument might lead students to *underestimate* (Beeckmans et al., 2001; Eyckmans, 2004; Mochida & Harrington, 2006; Stubbe, 2012) their vocabulary knowledge, even when they very rarely or almost never check non-words (Shillaw, 1996; Stubbe, Stewart, & Pritchard, 2010). In more cases, the tool causes *overestimation* when learners check “Yes” for either non-words (Eyckmans, 2004;

Mochida & Harrington, 2006) or real words they have no knowledge of even though they do not check any non-words (Pellicer-Sánchez & Schmitt, 2012), which is one of the most criticized side of this test (Meara, 2010; Read, 1993, 2000; Stubbe, 2014). In fact, the main reason for including a considerable proportion of pseudowords (imaginary words which resemble real words, e.g. *flort*) among real word items in the initial checklist formats that were used with L1 speakers (Zimmerman, Broder, Shaughnessy, & Underwood 1977; Anderson & Freebody, 1982) was being able to control and correct for guessing and therefore solve the problem of overestimation.

The first researchers to introduce the pseudowords to the field to see how properly new test design works were Meara and Buxton (1987). They started by developing a test incorporating 60/40 ratio of real words and pseudowords, respectively. Later, Meara and Jones (1988) and Meara (1992) developed computerised Yes/No tests based on the same design. However, the percentage of "non-words" (Read, 2007), "pseudowords" (Beeckmans et al., 2001), or "imaginary words" (Meara & Buxton, 1987) varied. In the following years, different versions of this test have been developed to use with diffent groups of learners (see Beeckmans et al., 2001; Kempe & MacWhinney, 1996). Recently, another important application of this test format has become a part of the European DIALANG project (Alderson, 2005; Alderson & Banerjee 2001; Alderson & Huhta, 2005), which is a computerized test battery including 14 European languages (http://www.dialang.org). The advantage of this framework is that according to their test scores, it is possible to give testees information about their lexical abilities and give them further most appropriate language tests (Eyckmans, 2004; Eyckmans et al., 2007). This tool is also claimed by Beeckmans et al. (2001) to allow us to provide the test takers with estimates of their receptive vocabulary size. The test has been further developed into some other computerized Lex family versions: X-Lex (Meara 2005a; Meara & Milton, 2005), which covers words up to the 5K frequency band

and is an indicator of students' "overall proficiency levels" (Meara, 2005b, p. 21); Y-Lex (Meara & Miralpeix, 2006), which is made up of sample words taken from the 6-10K word frequency bands, and is proper for more proficient learners; and AuralLex (A-Lex) (Milton & Hopkins, 2005), in which the same words as those used in the X- Lex are included although they are tested phonologically rather than orthographically.

In regard to scoring, because of the widespread use of pseudowords even in present-day versions of the Yes/No test, how to score the learners' final test results is subject to some criticism (Zhang, 2013). In these tests, with the inclusion of non-words, two possibilities arise for each item when marked as known, so four types of responses come out: "hit" (marking a real word); "false alarm" (FA) (marking a pseudoword); "miss" (not marking a real word); and "correct rejection" (not marking a pseudoword). These possibilities lead to the concern of how the false alarms will be treated during the scoring process. Eyckmans (2004) and Schmitt (2010) mention two common approaches to this matter: a) applying some adjustment formulas; b) deleting tests where the number of selected nonwords exceeds the maximum limit. In the literature, it has been proposed that in case of overestimation, the test scores are adjusted downwards by using four different formulas so as to better reflect examinees' actual vocabulary size. These are h - f (Anderson & Freebody, 1982), cfg (correction for guessing, Meara & Buxton, 1987), Δm (Delta m) (Meara, 1992; 2010) and Isdt (Huibregtse, Admiraal, & Meara, 2002). Several studies have been done to investigate these complex scoring formulas (Beeckmans et al., 2001; Eyckmans, 2004; Eyckmans et al., 2007; Huibregtse et al., 2002; Mochida & Harrington, 2006; Pellicer-Sánchez & Schmitt, 2012; Stubbe, 2012). Yet, the issues of whether or not they are equally accurate and effective for different testees or which provides the best adjustment remain unsettled within the field (Harsch & Hartig, 2015; Stubbe, 2012; Stubbe & Stewart, 2012). That is why Beeckmans et al. (2001) claim that the "Yes/No format in its current form does not meet the required standards in terms of reliability" (p. 272) and "suffers from a bias which cannot be handled by one of the correction methods while maintaining a sufficiently accurate measurement" (p. 272). Owing to such worries about the efficiency of such formulas, some researchers prefer the second approach; that is, they set a maximum of non-words and discard the data as unreliable if it exceeds this threshold. For example for Schmitt (2010) and Schmitt, Jiang, and Grabe's study (2011), the threshold is maximum three non-words out of a total of 30 (10%), so they suggest excluding the data with a higher number of non-words.

Because of such problems, there are some studies in which non-words are included in the test, but they are not included in scoring at all. Mochida and Harrington (2006), for instance, assess the Y/N test performance of their subjects as a predictor of their subsequent performance on the VLT and report that raw "hits" are the best indicator of the VLT scores. In his study, Alderson (2005) as well creates various scores, such as "simple total", "simple correction" and "raw hits", to treat pseudowords for the VSPT employed in DIALANG (p. 85). In the final variable, he simply ignores pseudowords and just gives credit for correctly identified real words. The high mean score (.82) and correlation with simple total (.84), achieved by raw hits show that a simple count of identified real words is a good indicator of vocabulary knowledge aspects ("Simple total" is the total of words correctly identified as either pseudowords or real words). It is also reported in the study done by Beeckmans et al. (2001) that the very high reliability calculated with raw scores (.91) appears to be artefactual, while the corrected (cfg) score decreases reliability to .85.

Regarding the inconsistent scores obtained through correction formulas, in the current study, no such formulas will be applied. Instead, Alderson's (2005) "raw hits" procedure will be employed. Therefore, the calculations will be limited to hit scores (90 real words) and pseudoword results will not be included so that the data will allow for direct comparisons with the other two measures.

2.2.2.2. The design features of the yes/no vocabulary test. This test format is said to be successful at measuring students' receptive vocabularies in a fairly reliable way (Meara, 2010) as it yields reliable results especially with low-level learners (Meara, 1996a). While some studies suggest that it is a more reliable tool than a multiple choice measure (Meara & Buxton, 1987), according to some others self-assessments are not generally as reliable as cloze tests at predicting students' future performance. However, they are regarded reliable in terms of showing that students do not know certain words (Heilman & Eskenazi 2008).

In the test, examinees are given lists of sample words taken from ten frequency bands (1K-10K). They are required to show that they have knowledge of a meaning of a tested item by merely putting a check mark in the box next to it or selecting either "yes" or "no"; therefore, the test is considered to be measuring the form-meaning link (Schmitt, 2008), as does the VLT. In Figure 5 below, a sample excerpt from a pen and paper test is illustrated.

1 ☐ obey	2 ☐ thirsty	3 ☐ nonagrate
4 ☐ expext	5 ☐ large	6 ☐ accident
7 ☐ common	8 ☐ shine	9 ☐ sadly

Figure 5 Sample items from a Yes/No test (Meara 2010:18).

Some researchers argue that the Y/N format measures learners' word knowledge on the basis of word form recognition (instead of meaning) (e.g. Elgort, 2013; Eyckmans, 2004; Milton, 2009) since test takers are just supposed to identify the words they believe they know. However, according to Schmitt (2010, 2014) and Pignot-Shahov (2012), the test should be regarded as a "meaning-recall item" due to the fact that although any kind of knowledge demonstration is not required, testees are actually tested on their knowledge of word meaning (Laufer & Nation 1999).

The test has some some reported merits and practicality, such as being easy for researchers to construct, administer, and score and for learners to respond. The simplicity of the task makes efficient use of examiners' as well as examinees' span of time and enables testing a great number of people and a large sample of a language within limited testing time, which is important to make reliable estimates of vocabulary size (see Anderson & Freebody, 1982; Beeckmans et al., 2001; Culligan, 2015; Huibregtse et al., 2002; Laufer & Aviad-Levitzky, 2017; Lemmouh, 2010; Pellicer-Sánchez & Schmitt, 2012). It is also a very practical test which sets minimal demand on students (Harrington & Carey, 2009), especially in terms of strategic knowledge (Eyckmans, 2004). According to Meara (1996a), unlike many other standard measures, the test works well across different proficiency levels and seems equally appropriate for beginner and advanced level second/foreign language learners. Meara (1993) even states that when used repeatedly, it is possible to measure learners' vocabulary growth and track at which rate new words are acquired. All these, in sum, make the Yes/No Test one of the best alternatives in the field of vocabulary research.

2.2.3. The vocabulary size test. Another very popular test of word knowledge is the VST, which was originally designed by Nation and Beglar (2007a: 9) "to provide a reliable, accurate, and comprehensive measure of a learner's vocabulary size from the 1st 1000 to the 14th 1000 word families of English" (p. 9). It is commonly used for measuring first as well as second language learners' written receptive vocabulary knowledge (Coxhead, Nation, & Sim, 2014; Nation, 2012b; Nation & Coxhead, 2014) in almost any ESL/EFL context (Beglar, 2010). Though the test was created to fill in the gaps of the VLT, according to Hashimoto (2016), besides compensating for the gaps, it goes far beyond the VLT including 4,000 more words.

The test is regarded important since it provides an overall vocabulary size estimate rather than just indicating the level of mastery of each frequency band (Stoeckel & Bennett, 2015; Schmitt, 2010). Nation and Coxhead (2014) and Coxhead et al. (2014) also state that this test can also measure the form-meaning link as well as concept knowledge to some extent. Related with its functions, Beglar (2010) validated the VST under a Rash framework, and thanks to its reported merits, the test has become a common application for research purposes as a number of studies show (e.g. Atkins, 2010; Bundgaard-Nielsen, Best, & Tyler, 2011; Elgort, 2011, 2013; Lin & Morrison, 2010; Mizumoto, 2011; Nguyen & Nation 2011; Uden, Schmitt, & Schmitt, 2014).

2.2.3.1. The historical development of the vocabulary size test. The VST first appeared in *The Language Teacher* (Nation & Beglar, 2007a) and was reproduced in some other books such as *Focus on Vocabulary* (Nation & Gu, 2007), *Teaching Vocabulary* (Nation, 2008) and *Researching Vocabulary: A Vocabulary Research Manual* (Schmitt, 2010). It can also be found on some websites (https://www.victoria.ac.nz/lals/about/staff/paul-nation; http://www.lextutor.ca; http://my.vocabularysize.com/select/test). Presently, the test exists in two monolingual versions written in English, both in pencil-and-paper and computerised formats, and some bilingual variants in Korean, Japanese, Mandarin, Russian, Vietnamese, Persian, and Spanish (for examples see Elgort, 2013; Karami, 2012; Nation & Coxhead, 2014; Nguyen & Nation, 2011; Wang & Du, 2014).

While writing the test items, the first 12 of the 14 sections of Nation's (2006) word lists, using word family range and frequency figures from the 10 million token spoken part of the BNC (the British National Corpus and Corpus of Contemporary American English) available in <lextutor.ca> were revised and both the monolingual and bilingual version test items (constituting 85 per cent of the test) were sampled from them (Nation & Beglar 2007a). The authors created this written test by basing it on a spoken corpus as they acknowledged that students learn words in an order closer to this spoken ordering.

2.2.3.2. The design features of the vocabulary size test. As mentioned before, there are two monolingual versions of the VST written in English: the original VST with 140 test items tests 1-14,000 (14K) and the more recent one with 100 items measures 1-20,000 (20K) frequency-ordered word families. The latter has A and B versions, which are considered by researchers such as Amirian et al. (2015) and Nation (2012b) to be parallel and identical.

In the VST, the tested items are presented to test-takers in the form of multiple-choice questions which has four choices, so as argued by researchers it obviously employs a recognition format (Coxhead et. al., 2014, 2015; Jordan, 2013; Nation, 2001; Ozturk, 2007; Read, 2000). For some researchers, it tests passive recognition of form (Levitzky-Aviad & Laufer 2013; Stewart, 2014) as examinees are just supposed to recognise the tested item given to them in the sentence stem and choose its best or associated meaning written in the form of translation or definition; for some, the VST is a meaning-recognition format test (Pignot-Shahov, 2012; Schmitt, 2010). Below is shown an example item from the test (see Figure 6).

pub: They went to the **pub**.
a. place where people drink and talk
b. place that looks after money
c. large building with many shops
d. building for swimming

Figure 6 A sample item from the VST (Schmitt, 2010: 293-306).

Since the test is in such a multiple-choice design, it offers a variety of benefits. For instance, it eases the scoring process, makes marking as much reliable and efficient as possible, allows a wide variety of content sampling, makes the test appropriate for testing learners from various language backgrounds, enhances reliability, controls item difficulty level, and allows participants to demonstrate their knowledge of every target item (Beglar, 2010; McLean et al., 2015a; Nation & Beglar, 2007a; Nation, 2012b). In addition, it is stated by Nation (2012b, no page number) and Coxhead et al. (2015: 127-128) that "at the item level, the test measures receptive knowledge of a written word form", while "at the test level, it provides an estimate of total vocabulary size…".

The VST is an increasingly popular test measuring knowledge of the tested word within very limited context. In each test item, the tested word appears in a short, non-defining sentence context which, according to Nation (2012b) and Coxhead et al. (2015), shows what part of speech the tested item is, limits and slightly cues the word meaning by offering an example of target word use. In addition, as much as possible, the definitions are constructed around higher frequency vocabulary than the tested word item.

There are some other issues to consider about the VST in general. One of them is that in order to answer any test items, it is not necessary to have full knowledge about the target word since the test allows students to activate and apply their partial knowledge. For example, when testing the item “azalea”, test-takers are not tested on their general knowledge about different plant species. Instead, they should only be able to say that it is a plant (Nation, 2012b, no page number). Related with this is a debate about whether to include “I don’t know” option among choices or not. According to Zhang (2013), the VST should encourage students to use partial knowledge. Having the same point of view, Nation (2012b) further states that “I don’t know” option, especially together with a penalty, might discourage informed guessing, which possibly draws on testees’ subconscious knowledge. Therefore, if the aim in the VST is to credit students’ partial knowledge use, the original format that does not have the “I don’t know” option is recommended.

Another issue related to the VST is about the application of the test. It is argued by researchers that learners should sit all the frequency levels (Coxhead et al., 2014; Karami, 2012; McLean, Kramer, & Stewart, 2015b; Nation, 2012b; Nguyen & Nation, 2011) since they might answer some low-frequency items as correct through blind guessing although they are not within their actual level of word knowledge, called ‘‘Slumdog Millionaire’’ effect (McLean et al., 2015b, p. 27; Nation, 2012a, no page numbers). Some lower frequency level words can also be known because of various reasons such as their being loan words or cognates, words related to hobbies and interests of learners, technical words that learners are familiar with, or words learners have met before (Nguyen & Nation, 2011). In addition, while

Stewart (2014) suggests that the test levels should be limited as the format involves a 25% chance of correct guessing, Beglar (2010) and Elgort (2013) recommend that examinees should not attempt more than two levels beyond their ability.

As for the administration of the VST, students should be allotted about 40 minutes for the 140-item test, and almost 30 minutes to complete the 100-item tests (Nation, 2012b), even in a computerized form (Nation, 2012a). It is also discussed in literature that though in the original test the test items are arranged in frequency order, such an order is not essential (see Nguyen & Nation, 2011). It may in fact be better to mix all the bands and guarantee a good spread of low frequency words throughout the test to prevent learners from giving up at the lower frequency bands and maintain engagement (Nation, 2012b; Nguyen & Nation, 2011).

While scoring the test, a student’s final score on the 140-item test is multiplied by 100 (see Amirian et al., 2015; Elgort, 2013; Lee, 2012; Nation, 2012b; Nguyen & Nation, 2011; Schmitt, 2010). On the more recent 100-item version, which targets the first 20,000 word family, it is multiplied by 200 (see Amirian et al., 2015; Nation, 2012b). In addition, there is no correction for guessing as it might distort and alter the estimates of vocabulary size (Coxhead et. al., 2015; Nation, 2012b).

On the other hand, the shared assumption existing in the literature is that bilingual VSTs are more sensitive measures. However, it was revealed by Amirian et al. (2015) that there is a close correlation between the 20000 bilingual version they developed for their study and the 20000 monolingual version as well as the 14000 bilingual one; therefore, both versions could safely be employed interchangeably in different contexts and for different purposes. Previous research also support that a bilingual VST can function as effectively as the monolingual version in distinguishing between learners from various proficiency groups because it usually returns lower scores as students proceed towards lower frequency bands where they have to deal with more difficult words (see Elgort, 2013; Karami, 2012; Nguyen

& Nation, 2011). In addition, the general tendency in bilingual tests is using the words implied by the definitions (i.e. the first language synonym which is a single word or phrase) as the correct option and distractors rather than translating each definition word by word into the first language (Levitzky-Aviad & Laufer, 2013; McLean et al., 2015a; Nation, 2012b; Nation & Coxhead, 2014). As a result, students get rid of the problem of not being able to understand definitions written in L2, which require good reading skills and grammatical knowledge (Stewart, 2014).

In short, taking Nation's (2012b) test specification document and Nguyen and Nation's (2011) findings and suggestions into account, in this study, the test-takers had to sit the whole test. Moreover, since there were participants from elementary level who had limited language abilities, it was believed that a bilingual format would be more appropriate.

2.3. Comparing and Contrasting the VLT, the Y/N Test and the VST

In literature, different specifications regarding vocabulary test designs exist and they help to better describe a particular test format. For instance, taking into account Read's (1993) four dimensions (simple to more complex designs; verifiable versus self-report responses; breadth versus depth of knowledge; isolation versus contextualization of test items), the three vocabulary tests discussed above can be described as follows: In relation to the first dimension, all of them are simple test formats since they ask examinees to indicate the correct answer rather than to perform complex tasks. In terms of response types, while the VLT and VST use the verifiable response format, the Yes/No measure uses self-report. Regarding Read's third dimension, all these tests measure vocabulary breadth rather than depth. In this respect, they do not measure vocabulary in the form of production, but rather its avalibility for productive use (Laufer & Nation, 1999), which also means these tests measure only comprehension. As far as the last dimension goes, the VLT and Yes/No test presents all target items out of context, whereas the VST presents them in very limited context, but without any clues for the correct response. Therefore, they are measures of lexical competence, not lexical performance (Ozturk, 2015). In addition, following the terminology offered by Read (2000: 9), the tests in question are *discrete*, *selective* and *context-independent* measures of vocabulary size since all of them take on the role of measuring vocabulary knowledge as an independent construct, focus on particular vocabulary items, and make test-takers produce correct responses without referring to any context. Read (2000) also states that the focus of discrete/selective/context-independent measures is on recognition and recall. Regarding the definitions of these constructs mentioned before, these measures can be regarded as examples of a recognition test that focuses on understanding the target word form or its meaning, rather than on the ability to recall and produce either one (see Table 1)

Table 1

Different Dimensions of Word Knowledge Measured by Vocabulary Tests

	simple format	more complex format	verifiable responses	self-report responses	breadth of knowledge	depth of knowledge	isolation of test items	contextualization of test items (very limited)
VLT	√	-	√	-	√	-	√	-
Y/N	√	-	-	√	√	-	√	-
VST	√	-	√	-	√	-	-	√
	lexical competence	lexical performance	recognition	recall	discrete	selective	context-independent	
VLT	√	-	√	-	√	√	√	
Y/N	√	-	√	-	√	√	√	
VST	√	-	√	-	√	√	√	

In addition, Brown (2007) specifies three main qualities for evaluating a test: practicality, reliability and validity. According to Bachman and Palmer (1996), test practicality involves material resources, human resources, and time for developing or scoring the test. Firstly, in terms of material resources, all three tests are very practical as they are all pencil-and-paper tests, which are easy to make copies of. In addition, they are suitable for on-line administration. Secondly, in terms of human resources, the VLT and VST are more difficult to construct compared to the Y/N checklist test, which is a simple list presenting decontextualised words. Nevertheless, in the VLT, the cluster format does not necessitate writing as many definitions as in the VST, whose multiple-choice format requires every test item to have a four-option answer. In addition, while it is necessary to write a short sentence without any context clues of definition for every VST item, the VLT does not have such a requirement. For test-takers, the Y/N test is a very practical measure. As declared by Read (1993) checklists are the simplest formats, where the only thing required from students is to read a number of isolated words and then mark it as known or not. Multiple-choice format of the VST and word-definition matching format of the VLT, though, necessitate more extensive reading. However, the VLT requires less reading than the VST as its clusters consist of individual words as well as short defining phrases. Lastly, regarding time, it is reported in the literature that the Y/N test is the fastest one to sit since it takes 10 minutes on average (Batista, 2014; Milton, 2009). On the other hand, the 50-cluster VLT takes about 30 minutes, while the 140-cluster VST takes about 40 minutes. In sum, among these tests, the Y/N test is the most practical; then comes the VLT with its distinct advantages over the VST, which is more time-consuming.

In terms of reliability, due to the fact that the Y/N test uses pseudo-words, it raises some reliability concerns (Read, 2000). There is empirical evidence that shows this format lacks reliability due to the unreliable effects of invented words on the test scores (Beeckmans et al., 2001; Cameron, 2002; Eyckmans, 2004; Meara, Lightbown, & Halter, 1994). The reason might be the possibility that a learner transforms an unknown word into a known one and marks it “yes”, a phenomenon called “mock” hits by Anderson and Freebody (1982). For example, in their study, they noticed that the nonword “sham” was interpreted by poor learners as “shame”, and this phenomenon inflated the test scores. Also, as argued by Batista (2014), it is easy to imagine that an individual may say yes (“I know this word”) to a nonword that resembles a real word in one test setting, and it is also possible for the same person to say no (“I don't know this word”) to the same item in another. However, in the other two tests, it is less likely for learners to give different answers to the same question since they must demonstrate their vocabulary knowledge by selecting real words only, which is a reason to favor the VLT and VST over the Y/N test in terms of their potential reliability.

As for validity, the Y/N test seems so problematic because of simply asking examinees to respond “yes” when they recognize a L2 written word form. In a self-report checklist format, it seems rather difficult to verify what this “yes” response means or to confirm actual knowledge, since it does not require explicitly demonstrated knowledge of correct word meaning (Beglar, 2010; Laufer & Aviad-Levitzky, 2017). In this case, as Eyckmans et al. (2007), Jordan (2013) and Meara (1996a) argue, a test-taker’s confidence in his/her lexical competence may play an important role while providing responses. For some participants, being familiar with tested words or being sure that they exist in the target language without any knowledge of actual meaning might be enough, while others (usually more experienced students) might respond referring to their knowledge of meaning and mark target items as known only if can they use them productively. It is also highly probable in such tests that some subjects may get confused and mistake a tested word for another one they have in their lexicons with a similar form. All these mean the validity of this measure might be in doubt to some extent (Huang, 2006) due to the fact that it does not satisfactorily discriminate between word meaning knowledge and mere familiarity, even though some researchers claim that the words recognized in a Y/N form are actually the ones that are known or used by test-takers (e.g. Cameron, 2002). In fact, some previous studies clearly show that dichotomizing the concept “knowing a word” appears to be a major problem in Y/N tests, since participants might possess different response styles in these tasks (Huibregtse et al., 2002). For example, Waring and Takaki (2003) reported a 7 per cent drop on scores from a recognition test to an L2 → L1translation test. In Eyckman’s (2004) study as well, the number of the items marked “Yes” but could not translated correctly was almost 30.58%. This means within the 82.4% “yes” responses to real word items, about one out of three seemed to be the result of defective self-assessment. This ratio reached a higher level (nearly one half) in Eyckman et al.’s (2007) study, where students were not able to translate almost half of the real words even though they had marked them as known in the preceding Y/N test. Such poor self- assessments inevitably lead us to the interpreation that the testees’ response styles can be related to the amount of trust they place in their Y/N Test responses, regarding not only the non-words but also the *hit* responses and that the Y/N format might not be equally appropriate for every type of learner. Another reason can be the fact that different test tasks mentioned here measure different constructs. The Y/N task addresses Nation's (2001) question “What does the word look like?”, while the translation task focuses on the question “What does the word mean?”. As a result, it can be said that though different tasks may target the same lexical items, they may in fact test different aspects of vocabulary knowledge beacuse of its multi-faceted nature as the studies by Laufer and Goldstein (2004) and Laufer et al. (2004) show.

On the other hand, the VLT and VST allow much more than passive recognition of the word form, as examinees must overtly establish the form-meaning link. This means both measures provide direct evidence that the target word and its meaning are actually known (Milton, 2009; Read, 2007). However, according to Nation and Beglar (2007a: 11), the VST is a bit more demanding test than the VLT. In the VST, the words provided as distractors should fit reasonably well into the stem, which means the expected response and each distractor often shares elements of possible meaning (Coxhead et al., 2015; Ozturk, 2016). Therefore, it becomes more difficult for the learner to choose the correct answer. In fact, to be successful at providing correct answers, examinees must possess a quite strong idea about what each tested item means (Beglar, 2010; Nation & Beglar, 2007a) as supported by the study carried out by Ozturk (2016). Based on the results of her studies, Ozturk posits that the two tests are possibly tapping receptive knowledge at different degrees. Since the distractors and the target word share elements of meaning, the VST requires more precision compared to the VLT, which aims at measuring partial lexical knowledge by using distractors that lack form-meaning relations to target words. To explain this better, she gives the example test item for the word *pub*, in which all options occupy the semantic feature of the word "location" and collocates with the verb *go*. In order to answer this test item correctly, test takers must more precisely know the kind of a place a pub refers to (see Figure 6). Sharing the same point of view as Ozturk, Batista (2014) believes that the VST can make better knowledge distinctions. Table 2 given below contains the summary of Brown's (2007) criteria and how the abovementioned three tests meet them.

Table 2

Summary of Characteristics of Three Vocabulary Tests - Brown's (2007) Qualities

	Practicality	Reliability	Validity
Levels Test	√√	√√√	√√√
Yes/No Test	√√√	√	√
Size Test	√	√√√	√√√

Note. √√√ = very high, √√ = high, √= not so high

The most important thing to highlight in this discussion is that, though both checklist tests and those which expect a person to link a word with a possible meaning, or vice versa, are all considered tests of lexical breadth, according to Milton and Fitzpatrick (2014:7), they may produce "quantitatively different estimates of vocabulary size given the different constructs which underlie the tests". The reason, as agreed by Schmitt (2010) is that size tests actually measure form-meaning knowledge to a different degree. For instance, the VLT and VST appear to be more valid reflections of students' recognition knowledge compared to the Y/N test as the Y/N test does not measure the form-meaning link explicitly. On the other hand, the underlying constructs of VLT and VST formats may refer to different vocabulary competence. Because of that, in spite of the general belief about either format that it can categorically measure recognition, it would be better to suggest that they should be regarded as different measures which target either receptive recognition or productive recognition with reference to the distinction drawn by Nation (2001). In this case the VLT can be claimed to be measuring a higher level word knowledge described in literature as active / form recognition and the VST that of passive / meaning recognition.

Another thing all these test formats have in common is that they are each typically guided by frequency measures (Read, 2007). Though frequency is in fact a continuous variable, the test items are drawn from lists where words are divided into 1K bands of frequency to make vocabulary testing more convenient (Meara, 2010; Ozturk, 2016). In fact, students’ knowledge of high or low frequency words is assumed to be the indicator of their overall vocabulary size as vocabulary knowledge seems to be strictly bound by frequency. That is why learners with smaller vocabulary size usually have insufficient knowledge of even higher frequency words, whereas those who have greater vocabulary size know relatively infrequent words and are more proficient in the target language (Huang, 2006). In this case, these tests are useful instruments to examine learners’ general language proficiency and academic achievement as suggested in the literature (e.g. Laufer et al., 2004). Therefore, the next section will be devoted to the relationship between frequency and vocabulary size, and vocabulary size and language proficiency, respectively.

2.4. Word Frequency Effect on Word Knowledge

It is argued in literature that frequency and vocabulary acquisition have a close association as frequency of usage determines acquisition for both native and non-native language learners (Daller, Milton, & Treffers-Daller, 2007; Ellis, 2002a, 2002b; Nation, 2006). High-frequency words are met more often in typical written texts and used in daily communication contexts most commonly (Nation, 2001; Schmitt & Schmitt, 2014). The effect of frequency is further enhanced in L2 teaching materials in which input is usually graded in terms of vocabulary difficulty. Therefore, these words are acquired earlier, more easily and usually known much better than lower frequency ones. In this case, it can be said that taking frequency into account, we can make a prediction about the probability of any lexical item being encountered in a given context but also the likelihood of its being recognized or known by learners (Alonso, 2013; Bennett & Stoeckel, 2013; Eyckmans, 2004; Read, 1988, 2000; Zhao & Ji, 2016). In Ellis' (2002a:144) terms frequency is "a key determinant of acquisition…", and for Brown (2012:20) it is recognized as "a key driver of all aspects of language learning, and indeed of human learning in general." Following Palmer's (1917: 123) assumption that "...the more frequently used words will be the more easily learnt.. . ", Meara (1992, 2010) suggested a model, known as the frequency model of vocabulary learning, for L2 learners' vocabulary development of size that can be tested empirically. According to his hypothesis, as learners' proficiency develops, first they will almost fully acquire the most frequent 1,000 words and continue to develop a quite accurate knowledge of the next 1,000 words. Yet, a gradual decrease will show up in the third and fourth thousand word frequency bands, and a flattening of the slope will appear beyond the 5000-word frequency level (see also Milton's "frequency model of lexical learning" (2007:48), formulated to show that frequency has strong effects on L2 vocabulary development and supported by his own research (2007, 2009)). As shown in Figure 7, with a decrease in the word frequencies, a typical learners' familiarity with the lower-frequency words will be increasingly less.

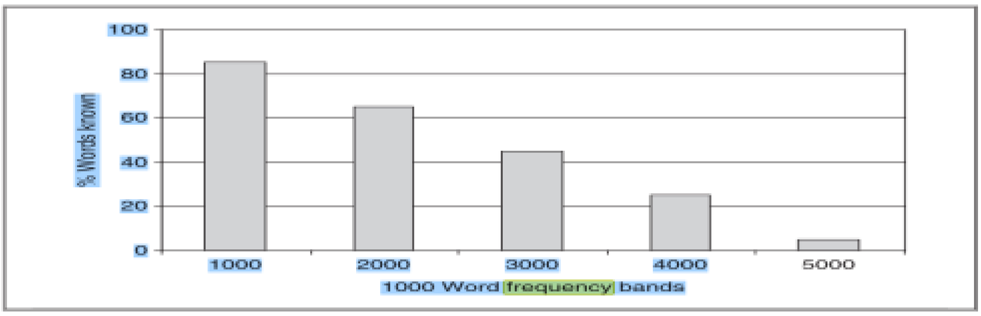


Figure 7 Vocabulary profile of a typical learner (Meara, 1992: 4)

Regarding the notion that frequency level of a word and the probability of its being known are highly interrelated, vocabulary size measures are naturally based on frequency. In these measures, words are grouped by frequency, and the underlying assumption is that test items at differing frequency levels would form a difficulty continuum. If this assumption is in fact plausible, then examinees' scores should decrease when they near less frequent bands.

In literature, there are some studies whereby the frequency model is confirmed. For example, Aizawa's (2006 as cited in Brown, 2012) study covered the eight 1000-item levels of the JACET8000 list, and including Japanese university students, he tested how many items are known. The results showed that the frequency model was rather effective in the higher frequency levels as overall scores decline gradually over the first four bands. Beyond this level, however, some other factors appear to be effective. Similarly, giving the X_Lex to 227 participants, Milton (2007) looked at their vocabulary profiles. The overall results showed that the frequency model worked well (see also Milton, 2009). Brown (2012) carried out a small-scale partial replication of Milton's work and tried to find out how well the frequency model applies to a group of Japanese university students. The group results obtained from a Yes/No test revealed the same expected pattern as it was clearly understood from the scores that the tested items in the highest frequency band were best known and the ones in each subsequent level were less well known. There are some previously conducted studies that provided evidence for varying degrees of learning rate between different frequency levels and supported the frequency-based developmental pattern (e.g. Laufer et al., 2004; Laufer & Paribakht, 1998; Milton, 2007, 2009; Ozturk, 2016, 2015; Richards, Malvern, & Graham, 2008). It was revealed in these studies that gaining mastery of vocabulary follows a declining order starting from the highest frequency bands and heading towards the lowest ones as there was a linear decrease in scores over frequency levels. Then it can be said that all the abovementioned studies back up Milton's (2007:57) conclusion that "the frequency effect on vocabulary learning is very strong and this should not be lost”. Milton also posits that vocabulary size and knowledge measures based on word frequency retain extremely good construct validity.

There are also some studies in literature which reveal that the frequency bands of some size tests form an implicational scale, whereas there are also some which do not support the presence of such an implicational scale. For example, in three studies where the vocabulary measure was the receptive VLT, an implicational scale was established (Ozturk, 2015; Read, 1988; Schmitt et al., 2001). However, in two studies conducted by Ozturk (2016, 2013) with different types of L2 learners, the employed measure was the VST, and though learners’s scores decreased with frequency levels, an implicational scale could not be formed. According to the author, the absence of an implicational scale in the VST scores was not expected since it is as well a measure that is able to tap into similar knowledge type in learners who have almost the same language abilities. In fact, the results obtained from the abovementioned studies show that these tests cannot be considered as equavalents in terms of difficulty and support Nation and Beglar’s (2007a:11) claim that the VST is slightly more demanding than the VLT, which requires L2 learners to have a less precise idea of the target word meaning and form. Based on such findings, Ozturk (2016:13) reports that “The two tests are likely to be tapping different degrees of receptive knowledge”.

On the other hand, the findings of the studies carried out by Elgort (2013), Karami (2012), Nguyen and Nation (2011) and Zhao and Ji (2016) supported the frequency effect on bilingual versions of the VST. In these studies, learners’ scores indicated a linear or rough decrease among frequency levels and suggested that words of higher frequency levels were mastered faster than those of lower frequency bands.

Another point related to the vocabulary size measures is that this frequency factor becomes more problematic especially in lower frequency bands. It has actually been indicated by Milton (2009) that as implied by Zipf's Law, the effect of word frequency diminishes in less frequent levels and the word knowledge of an individual learner rapidly diverges (Greidanus & Nienhuis, 2001) due to various reasons, such as the chance factor (McLean et al., 2015b) or the fact that lower-frequency words are more specific to particular language genres (Li & MacGregor, 2010). Therefore, in literature it is usually advised that vocabulary knowledge testing should be limited to 5000 word frequency band as that many words are considered to be within the learning goal for non-native university students, regardless of their fields of study (Cameron, 2002; Hazenberg & Hultsijn, 1996; Hsueh-chao & Nation, 2000; Laufer, 1998; Nation, 1990, 2006; Webb & Sasao, 2013).

Based on the discussion so far, the main focus of the present study will be on the power of the three frequency based size tests to work similarly and distinguish between different proficiency levels; measure the same single underlying construct, presumably passive knowledge of vocabulary at recognition level; and provide the same results at different frequency levels. In this respect, the participants should perform better on the higher frequency words and get higher scores from band 1 than from band 2 (1000 < 2000), while the scores received from band 2 should be equally high or higher than those on the following sections. Also, regarding the fact that students' word knowledge predominates in higher frequency bands and that less frequent words are heavily dependent on their particular needs or interests and following Read's (2007) advise that vocabulary size measures for groups of second language learners should target narrower ranges of word items as low frequency words are far less likely to be known in foreign language environment, only the first five frequency bands (5000 word family) were tested in the present study. It was also thought that limiting the frequency sections would "prevent learners from getting frustrated with too many unknown words in the lower frequency levels" as stated by Ozturk (2013:8).

2.5. Vocabulary Size Effect on Language Proficiency

It is widely recognized that knowledge of vocabulary has close ties with overall language competence. In general, vocabulary supplies the students with the “enabling knowledge” that is necessary to become successful in other areas of language proficiency (Laufer & Nation, 1999). Due to its contribution to all other language skills, vocabulary knowledge is acknowledged as a reliable proxy for general foreign language proficiency (Elmasry, 2012; Koda, 1996; Milton, 2013, 2009) or a benchmark of proficiency in both receptive and productive skills (Maskor & Baharudin, 2016). Gyllstad (2007) also highlights the importance of vocabulary size as an essential factor and a rough indication of general proficiency, as confirmed by ample studies in which vocabulary size test scores tend to correlate well with scores obtained from different tests measuring various language skills (Alavi & Akbarian, 2012; Albrechtsen, Haastrup & Henriksen, 2008; Golkar & Yamini, 2007; Laufer & Ravenhorst-Kalovski, 2010; Milton & Alexiou, 2009; Milton, Wade & Hopkins 2010; Nation & Meara, 2010; Saville-Troike, 1984; Stæhr, 2009; Tseng & Schmitt, 2008)

> “What [the DIALANG analysis] would appear to show is that the size of one’s vocabulary is relevant to one’s performance on any language test, in other words, that language ability is to quite a large extent a function of vocabulary size.” (Alderson, 2005:88)

In Table 3 given on the next page, some previous studies in which learners’ knowledge of vocabulary measured by different vocabulary size tests and their scores at different tests covering various langugae skills were presented. Taking the findings of many vocabulary related studies, we can say without doubt that there exists a very strong correlation between vocabulary size and language proficiency.

Table 3

Previous Studies Employing Different Vocabulary Size Tests

Author(s)	Participants	The Tests Correlated	Correlation score
Beglar and Hunt (1999)	496 learners	the first two levels of the VLT and TOEFL scores	a high correlation between the two measures (r = .70)
Golkar and Yamini (2007)	76 Iranian undergraduate students	the receptive and productive versions of the VLT and the TOEFL test scores	passive vocabulary knowledge correlates well with language proficiency (r = .91)
Milton and Alexiou (2009)	575 SL learners from various language learning contexts	scores of X_Lex against the Common European Framework of Reference for Languages (CEFR) levels in French and Greek as foreign languages	in Spain & Greece, the CEFR level attained seems particularly sensitive to learners' vocabulary knowledge, and 60-70% of variance can be explained by vocabulary size; in Britain, a strong relationship is again observed, so over 40% of variance can be explained likewise
Milton, Wade, and Hopkins (2010)	30 participants	International English Language Testing System (IELTS) scores and X_Lex (Meara & Milton, 2003) / ALex (Milton & Hopkins, 2005) reading (R), listening (L), writing (W), and speaking (Sp)	X_Lex correlated well with overall IELTS scores (r=.68) and A_Lex (r=.55) strong correlations between written and aural receptive vocabularies as well as vocabulary size and L2 language ability
Staehr (2008)	88 testees	exam grades on receptive skills as well as writing papers and the testees' scores on the receptive VLT (Schmitt et al., 2001)	a strong link between vocabulary knowledge and reading (r =.83): about 72% of variance in reading can be explained by vocabulary size; but, not such a strong link between vocabulary knowledge and writing (r =.73) or vocabulary knowledge and listening (r =.69): almost 52% of variance in writing, and 39% of variance in listening can be explained by word knowledge

Staehr (2009)	115 advanced Danish EFL learners	a standardized listening test from the Cambridge certificate of proficiency in English (CPE) (2002), VLT (Schmitt et al., 2001) and Depth of Vocabulary Knowledge Test (DVK)	a significant correlation between vocabulary size and listening comprehension r =.70
Alderson (2005)	reading: 718 grammar: 1084 writing: 735 listening: 606 vocabulay: 975	Vocabulary Size Placement Test (VSPT) in the DIALANG project and language tests measuring five macro skills: reading (R), grammar (GR), writing (W), listening (L) and vocabulay (V)	raw hits (a simple count of real words correctly identified) vocabulary has a strong link with four language skills: reading .53, listening .52, writing .59, grammar .54 and vocabulary .63. the results indicate that vocabulary knowledge accounts for 37–62% of the variance in the other language proficiency scores
Meara and Jones (1988)	109 learners at the Cambridge Eurocentre School, and 159 learners in London	VOC (the vocabulary test) and the Eurocentres JET test	Cambridge: overall correlation: .664 London: overall correlation: .717

Enayat and Amirian (2016)	135 EFL learners	VLT and WAT (Word Associates Test) and OPT (the Oxford Placement Test)	no significant correlation between WAT and OPT,
			r = -.04, p > .05.
			a significant correlation between VLT and OPT,
			r = .39, p < .01.

Stubbe et al. (2010)	97 lower level Japanese university students	Y/N & a bilingual vocabulary test with the same words in the format of the VLT to check for overestimation	VLT scores 70.9% substantially higher than Y/N scores 50.7%
Stubbe, (2012)	490 Japanese university students (5 higher proficiency & lower ability universities)	a VLT style MC test with 96 real words from YN forms, plus three extra words	contrary to some similar studies (e.g. Mochida & Harrington, 2006), the participants tended to underestimate their sizes of vocabulary on checklists

The findings of these studies also imply that the correlation evidence between the vocabulary size tests and other measures of general language proficiency must not be regarded as a casual relationship. In fact, it is stated by Milton (2009) that size tests provide the same information as other language proficiency measures, and for Laufer et al. (2004) provided that the correlation level between vocabulary size and language skills is high, a vocabulary size test might be a useful tool to measure proficiency. For example, they found in their study that knowledge of form-meaning link accounted for 42.6% of the total variance in their participants' class grades. It is also clear that while many researchers consider the VLT to be a reliable and valid measure of vocabulary size (e.g., Beglar & Hunt, 1999; Cameron, 2002; Qian, 1999, 2002; Schmitt et al., 2001) and a predictor of language proficiency (Enayat & Amirian, 2016), some favor the Y/N test due to high correlations between it and other global proficiency tests (Eyckmans, 2004) and advise using it either at the beginning of a course to place students in an appropriate group or at the end to determine their vocabulary sizes (Nation, 2000).

In short, with the conclusion that there is a close correlation between vocabulary size and overall language proficiency and that vocabulary size tests can be used as predictors of language skill levels as well as L2 proficiency, the next question naturally arises: Which of the three vocabulary size measures used in the present study provides a score that correlates best with participants' university preparatory class exit scores? If results can indicate that any one of these three tests is able to predict performance on the proficiency exam, it can be offered as an alternative to the proficiency exam used in the institution where the researcher works to measure students' language proficiency. Regarding the suggestion made by Stæhr (2008) that "in order to explore fully the relationship between vocabulary size and language proficiency, future studies need to use multiple measures of vocabulary size" (p. 148) and the fact that there are no empirical studies correlating the abovementioned variables, this will be in the focus of the current study.

2.6. Studies Comparing Vocabulary Size Tests

In literature, there are numerous validation studies that compared the results of Yes/No tests with those of other vocabulary size tests, primarily the ones using a multiple-choice format. Based on their findings, it can be said that though these studies traced correlations between the test instruments, just some of them found strong correlations and supported the concurrent validity of the Y/N test ($r \geq .50$) (e.g. Anderson & Freebody, 1982; Harrington & Carey, 2009; Lemhöfer & Broersma, 2009; Meara & Buxton, 1987; Meara & Jones, 1988; Mochida & Harrington, 2006), while some others did not (e.g Cameron, 2002; Stubbe et al., 2010).

For example, in the validation study carried out by Anderson and Freebody (1982), 120 fifth graders, all native speakers of English, completed a Yes/No Test consisting of 195 real words as well as 131 nonwords and a multiple choice test involving the same word items. The correlation between the scores of the multiple choice and corrected Yes/No was .84. According to the authors, considering the fact that the same 195 words were assessed in both tests, this value was considerably low and did not represent a strong relationship. In addition, since these two formats do not measure exactly the same thing, the main question to which they were seeking an answer was which test offers the most valid vocabulary knowledge assessment. To find this out, they interviewed the participants by making them read a target word and either define or use it in a sentence. The interview scores correlated much better with the scores of the Yes/No format than those of the multiple choice which led the authors to conclude that "[…] a score on a yes/no test provides a much more valid indicator of whether an examinee actually knows the meaning of the tested word than a score on a standardized multiple choice test" (p. 37). On the other hand, the authors also agree that the

Yes/No test may have lower "reliability" and "predictive validity". The basis they argue for such caveat is that in order to perform successfully on a multiple choice or matching test besides word meaning knowledge, students should have the ability of reasoning, using working memory in a planful way to help hold response options in mind and understanding subtle nuances between L1 and L2. Such skills and knowledge are usually possessed by learners of high ability or socio-economic status and apparently contribute to reliability and predictive validity. However, in their study, Anderson and Freebody’s young and underachieving participants who were less likely to possess strategic knowledge or test wiseness did not possibly pay attention to the nature of distractors or consider all options but rather picked the first one that stroke their fancy, which might have affected the test scores. In other words, although the Yes/No Test seemed to provide better measures of word knowledge than the multiple choice test with young learners, the same study might have given totally different scores with older students. It should also not be forgotten that the inclusion of nonwords in a Yes/No test is also a factor that might affect the reliability of the test in a negative way as mentioned before.

This study led to further research into both tests for their efficacy in determining L2 learners’ vocabulary size. For example, Meara and Buxton (1987) firstly compared a specially constructed Yes/No test with a multiple-choice test similar to Cambridge First Certificate Examination (FCE) in terms of design to determine the relationship between them, and secondly they compared the test scores of some subjects with their results of the FCE to see which test is better at predicting the non-native speakers’ exam grades. 100 subjects took the first two tests and 26 of the 100 subjects took the FCE. According to the authors, the correlation (around r =.7) between the two measures was satisfactory and indicated that in spite of the evident differences between them, the MC and Y/N tools were measuring mostly the same kind of thing. However, when the participants’ test scores were correlated with their examination grades, only the Yes/No figures were reported to be significant (χ^2=13.6, p < .01 with 2df) regardless of the fact that in that exam, a multiple choice vocabulary task was actually the major component. The main difference between this study and the one by Anderson and Freebody is that in the former, the tested items were not identical, and the subjects were not a homogeneous group. In addition, the participants took both tests on the same day, with a 15-minute break.

Cameron (2002) also conducted a research with secondary ESL students in the UK. She assessed the VLT (Nation, 1990) to see whether it is an adequate test to use with younger learners of English as an additional language by comparing the students' performance on this test and the Yes/No. In contrast to some earlier studies, the participants' performance on these measures did not correlate. Correlations between the scores obtained at three frequency levels of the test as well as the Academic Word level ranged from .15 to .45, and none of them reached a significant value. Above all, though identical items were covered at the 2K and AW levels in both tests, the correlations between them were approximately .20.With this study, Cameron shows that in an additional language learning context, the VLT offers a much more adequate tool compared to a more practical Yes/No test, and the tests do not seem interchangeable due to the fact that the VLT is somewhat more demanding.

In the following years, in order to be able to ensure concurrent validity for the Y/N measure, Mochida and Harrington (2006) examined a group of students' performance on the Y/N (Huibregtse et al., 2002) against that of the VLT (Nation, 1990). Their participants were 36 undergraduate and postgraduate students. Different from Meara and Buxton (1987) and Cameron (2002), they compared the examinees' performance on the same target words across all levels of frequency, as in Eyckmans (2004), and found that the two measures correlated at r = .83. In this study, Mochida and Harrington examined the four correction-for-guessing formulae along with raw hits scores. Whereas correction formulae were reported to have a tendency to increase the difference between the scores on both measures, the number of raw hits, that is, the “Yes” responses to real words only was stated to be the best predictor for students’ overall performance on the VLT. Another finding was that the performance on the Y/N test accounted for over 75% of the variance in the overall VLT scores. This means the Y/N test has the power to measure the same knowledge type as correctly as it is done by the VLT. This cross-validation study in fact shows plenty of construct validity of a Yes/No test. Based on these results, the authors conclude that the Y/N measure can serve as a practical and useful alternative to the VLT, and presumably, other similar formats.

In a recent study by Stubbe et al. (2010), scores of the participants on the Yes/No measures were compared to their subsequent scores on a bilingual vocabulary test including the same words in the format of the VLT particulary to determine the potential effects of vocabulary overestimation. Interestingly, the scores on the VLT-format test were reported to be substantially higher, with a mean of 70.9%, compared to the mean of 50.7% on the Yes-No instruments (N = 97). According to the results, the authors concluded that contrary to some similar studies (e.g. Mochida & Harrington, 2006), the participants included in this experiment, who were lower level Japanese university students, tended to underestimate their sizes of vocabulary on checklists.

In another study with a similar approach, Culligan (2015) employed the Y/N and VLT and found that the correlation between these measures was was at .63. In this study, for both tools, the difficulty parameters were estimated through a one-parameter response model for dichotomous data. The Y/N scores represented only the Hits, without any alternative correction formulas, and the VLT scores were composed of the total number of the test items that were answered correctly. The author considers the correlation as strong, and according to him, test scores provide evidence not only for concurrent validity of the Y/N test but also for the interpretation of test scores as indicative of vocabulary knowledge.

So far, the discussion has been based on the relationship between the Y/N Test and the MC and the Y/N Test and the VLT and the recent hypothesis that the form-meaning link can be operationalised in different ways. One obvious thing here is the fact that though size and strength are related constructs, different knowledge types are expected to cause different test modes to produce a variety of results, as they vary in strength. Therefore, following the idea of testing the form-meaning association in a variety of test modalities, the goal of this study is to test the participants' receptive vocabulary, and specifically to examine whether each size test will provide the same score as they are considered to be measuring the same construct underlying the same strength modality. In order to do this, the form-meaning link will be measured through three different passive recognition size tests. In the VST, the participants will be given target word forms and they will be supposed to recognise their meanings from the given options; in the VLT, the meanings of the tested words will be presented to the participants, and this time they will be asked to recognize and select their matching word forms from among the given options; and in the Y/N Test, the participants will only see groups of listed items and they will have to recall the meaning of each without any requirements for production to be able to supply the expected response. In fact, it is obvious in the literature that there are some other tests that can be used to measure vocabulary knowledge; however, these three test formats are the most commonly used ones. Also, although productive dimension is acknowledged to be extremely important, recognizing word form is regarded as the most basic word knowledge level. Moreover, although a word recognition test can only tap into a small extent of the complex network of the overall vocabulary knowledge of the language learner, a word recognition sum can actually be a worthwhile indication of the learner's vocabulary knowledge outer limits as recognizing a word is a precondition to understand it or to use it with any depth of meaning. On the other hand, in some earlier studies, in order to provide convincing evidence of concurrent validity for the Y/N Test, VLT and VST, content and format were usually confounded since researchers did not use identical items (e.g. Meara & Buxton, 1987; Shillaw, 1996, and most items used in Cameron, 2002), and those whereby identical items were tested were limited to the comparison of performance on only two tests, mostly the Y/N measure and VLT (e.g. Eyckmans, 2004). Here, the participants' performance on three size tests with the same target items and a similar design feature to the abovementioned three tests will be examined. This direct comparison of multiple assessment performance will provide better evidence for the construct validity of the tests, which lacks in the field of L2 vocabulary research. Unlike the decontextualized Y/N format, the VLT and VST are cued recognition tasks which are expected to result in higher performance. What is more, since the three measures employed in this study are not considered to be the same concerning difficulty, the main focus will be on the questions to what degree test scores will differ from one another and whose score will provide the highest correlation with the participants' university preparatory class exit scores. In this respect, the answers to the following research questions will be sought:

1. Do the three English vocabulary tests, the Vocabulary Size Test (Nation & Beglar, 2007a, 2007b), Vocabulary Levels Test (Nation, 2001) and Yes/No Test (Meara, 1992), reveal similar overall receptive vocabulary knowledge estimates in different proficiency levels?

2. Do the VST, VLT and Yes/No Test reveal similar overall receptive vocabulary knowledge estimates across test sections?

3. Do the VST, VLT and Yes/No Test reveal similar overall receptive vocabulary knowledge estimates in different frequency bands for different proficiency levels groups?

4. a. How well do the VST, VLT and Yes/No Test correlate with one another?

b. Which of the VST, VLT and Yes/No Test has the best correlation with the participants' university preparatory class exit scores and yields consistent results?

Chapter III

Methodology

This part explains the research design and methodology of the study. It presents the method for carrying out the study by introducing the participants in the selected setting and describing the instruments and procedures for data collection. It follows with the report of the pilot experiment in which the appropriateness of the materials was tried to establish and finishes with the administration of the instruments.

3.1. Research Design

In the present study, since quantitative aspects of vocabulary knowledge like size and strength were targeted, the research design was based on quantitative assessment of vocabulary. Employing quantitative procedures would offer the advantages of testing more words and using multiple assessments, which in turn would provide better construct-related validity evidence.

3.2. Participants

In order to gather data for the current study, convenience sampling (a type of non-probability sampling in which the researcher gets help from the available subjects (Fatemipour & Jafari, 2015)) was used. The participants of the study consisted of 581 learners studying at the English preparatory classes of two different state universities in Bursa. The medium of instruction in those universities is both English and Turkish. There were 316 male and 265 female participants. They ranged in age from 18 to 22. English is their foreign language. At both institutions, at the beginning of the academic year, students are given an English Proficiency Examination, and those getting 70 and higher grades on this exam start their education in their departments. The students whose English is not sufficient enough to pass this exam are divided into levels according to the result of the placement test and start to study in English Preparatory Class for a year in groups of 25 to 30. The classes in every level are formed randomly and include students who are going to start their undergraduate courses in different majors, which means they are not homogeneous with respect to major but similar with regard to their linguistic proficiency levels. Taking into consideration the participants' proficiency exam results that they took at the beginning of the academic year, the students from three different levels (Intermediate, Pre-intermediate, and Elementary) were chosen for this study. In Table 4 given on the next page, you can see in detail how many students from each level were involved in the present study.

Table 4

The Levels and Total Numbers of the Participants

	The 1st State University		The 2nd State University		
	Male	Female	Male	Female	Total
Elementary	75	94	31	11	211
Pre-intermediate	59	64	88	47	258
Intermediate	57	44	6	5	112

When the students were given the tests, they were studying A1-A2 materials in elementary level; A2-B1 materials in pre-intermediate level, and B1-B2 materials in intermediate level at both institutions accordingly with Common European Framework of Reference for Languages (CEFR). They also had supplementary packs for grammar, reading, writing and vocabulary classes which were compiled from different (re)sources in parallel with the topics covered in each course. For example while the grammar pack included to the point exercises for extra practice, the reading and vocabulary packs provided the students with some academic texts and vocabulary which were related to their majors. The writing pack, on the other hand, presented a variety of paragraphs and essay types for further writing practice. Students had a skills-based program in which they had five different courses in all levels at the first institution: Grammar, Reading, Writing, Listening & Speaking and Vocabulary. At the second institution, the students followed a similar intensive teaching program in which there was a Main Course class along with Reading, Writing, Communication and Integrated Skills classes. In both institutions, vocabulary is thought and tested separately or inclusive into the other skills such as listening or reading. Tables 5 and 6 show which classes and how many hours a week students are expected to take each class in both universities.

Table 5

The Classes and Weekly Hours of Each Class Offered in the 1st State University

Classes	Weekly Hours of Each Class at Different Proficiency Levels		
	Elementary Level	Pre-Intermediate Level	Intermediate Level
Grammar	7	5	4
Reading	5	5	5
Writing	5	5	4
Listening&Speaking	7	7	7
Vocabulary	2	2	2
Total hours	26	24	22

Note: + video projects, two writing evaluations, and two extensive reading quizzes each term constitute 20% of the students total grades.

Table 6

The Classes and Weekly Hours of Each Class Offered in the 2nd State University

Classes	Weekly Hours of Each Class at Different Proficiency Levels		
	Elementary Level	Pre-Intermediate Level	Intermediate Level
Main Course	18 (A2)	18 (B1)	
Reading	4 (A1+A2) Extensive Reading Supplementary Material	4 (A2+B1) Extensive Reading Supplementary Material	6 (B1+B2) Online Reading (Read Theory)
Writing	4 (A1 Book)	4 (A1+A2 Book)	6 (B1 Book)
Communication	4 (A1)	4 (A2)	12 (B1)
Integrated Skills			6 (B1 + B2)
Total hours	30	30	30

3.3. Data Collection Tools

3.3.1. The target vocabulary. Three adapted bilingual (English-Turkish) vocabulary tests with the same desing features as the Vocabulary Size Test (Nation & Beglar, 2007a, 2007b), the Vocabulary Levels Test (Nation, 2001) and the Yes/No Test (Meara, 1992) were used in the present study. The reason for not using the original monolingual tests was to be able to test only the construct of vocabulary knowledge rather than reading skills or knowledge of complex grammatical structures that are required to understand definitions given in the VLT and distractors used in the VST (Nguyen & Nation, 2011). Such factors not only make the monolingual tests more challenging and time-consuming but also contaminate the measurement (Karami, 2012). It is also stated by Elgort (2013) that bilingual tests can reduce anxiety and examinee fatigue; therefore, they are expected to allow for more accurate estimation. In order to avoid such problems, the participants were presented with Turkish equivalent of an English target word rather than word by word translation of its definition into the learners' native language. The short bilingual definitions or distractors would probably prevent the participants, especially the low-level ones with insufficient language capabilities, from the burden of trying to read and understand much longer monolingual ones which require not only good reading skills but also some cognitive abilities.

On the other hand, all the three tests used in this study were receptive recognition tests measuring the same 90 target words which in total represent most frequent 5000 word families (5k) in English (see Appendix 3). As understood, there were 18 target word items selected as a representative of each frequency level - nine nouns, six verbs, and three adjectives. There are two important reasons here to limit the tests to 5000 word families band and exclude the 10000 word families band (10k). Firstly, 10k is thought to be far beyond the language proficiency of the participants. Secondly and most importantly, the gap between 5000 and 10000 words bands would make the cross-validation of the test scores rather confusing as the 5000 words between these bands are measured within just one section in the VLT, though they are spaced evenly and measured within separate sections in the VST.

Item specifications that were followed while redesigning the tests (see Appendix 2) were reverse engineered from previous test descriptions (e.g. Nation, 2012b; Nation & Beglar, 2007a), and while retrofitting items taken from the three monolingual VST tests, specification-driven test assembly was implemented as recommended in Fulcher and Davidson (2007). Within the process of retrofitting and redesigning, 74 target items of the total 90, along with their distractors, were compiled from three different versions of the Vocabulary Size Test: 20,000-word monolingual test (versions A&B) and 14,000-word monolingual test (Accessed at <victoria.ac.nz/lals/about/staff/paul-nation>; http://jalt-publications.org/tlt/resources/2007/0707a.pdf; http://my.vocabularysize.com; http://www.lextutor.ca/). The remaining 16 items were selected randomly on the basis of frequency from word frequency lists based on the BNC/COCA corpus 1-25k. These additional test items were “bottle, danger, plant, delicious” (1K); “blind, current” (2K); “immigrate, disguise, envy, flock, beneficial” (4K); “congest, dubious, interrogate, applaud, versatile” (5K). Each was again checked in terms of its appropriateness to the correct frequency level of the BNC/COCA lists. On the other hand, the Yes-No Test (Meara, 1992) was used only as a source from which 30 non-words were selected (available at <lextutor.ca>). 61 of the abovementioned 74 items were also included in the New VLT (McLean & Kramer, 2015), which is itself adapted from the VST. Only the target word “abandon” was taken from the VLT, 2k-10k (Schmitt et al., 2001) (see Table 7).

Table 7

The Sources Used for Item Selection

90 items representing most frequent 1-5K & 30 non-words	
73 items	VST: 20K (A & B) & 14K monolingual tests (Nation & Beglar, 2007)
1 item	VLT, 2k-10k (Schmitt et al., 2001)
16 items	BNC/COCA corpus 1-25K (random selection)
30 non-words	Y/N Test (Meara, 1992)

In addition, all of the target items were cross-checked from the BNC-COCA 1-25k lists to ensure that they are used in their appropriate frequency band within the test. For example, some items such as *basis* was in the first 1000-word level of Nation’s VST; yet, it was re-assigned to the second 1000-word level and some other items like *nil*, which was presented in the second 1000-word level of the VST, are not used in this study as they do not take place in the first five 1000-word frequency levels of the BNC/COCA lists. The reason for this discrepancy might be that the new BNC/COCA lists, as stated by Webb and Sasao (2013), act as "representative of current English and provide a far better indication of the vocabulary being used by native speakers today than the lists used for the creation of the earlier versions of the VLT" (p. 267). What is more, during the test creation process, *Oxford Advanced Learner's Dictionary-8th Edition*, *Oxford WordpowerDictionary: English-English-Turkish*, Turkish English Dictionary: Tureng and Zargan English-Turkish Dictionary were used as well to check the correct and/or the most common usage and Turkish equivalent of every target word and its distractors.

As for the distractors, they were selected from among the words which belonged to the same frequency band as the target word, which means their difficulty level was the same as the target word, and they were used in the form of one- or two-word L1 definitions. In the current study, approximately 90 per cent of the distractors were different from the ones used in the original size tests as they were written by the researcher herself with the guidance and help of her supervisor. While writing the distractors, if the tested item was a noun, all the distractors were also nouns; if the tested item was a verb, all the distractors were verbs, which means the distractors and the tested item always shared the same part of speech. In addition, in the VST, while all the distractors were almost equally plausible within the context sentence, in the VLT, they were not. In order to agree upon the best alternatives, distractor analysis for each item in each test was done and the problematic distractors went through a continuous modification and editing process in accordance with the tremendous amount of feedback received from the researcher's former supervisor and her native and non-native colleagues. When the test creation process was over, first the tests were proofread by the researcher and her former supervisor. Then the final versions were proofread by three non-native speakers of English, two bilingual teachers and four native speakers for correct spelling, punctuation and grammatical errors and also to make sure that all the items on the tests have the correct usage and the very best definitions and distractors. The item whose own usage, translation or distractor usage was considered to be incorrect by the proofreaders was modified again. Due to such details, redesigning the bilingual formats of the tests took approximately six months.

Besides all these, the selection of the test items was modified along the following constraints as Eyckmans et al. (2007) did:

a) The whole test sample was restricted to nouns, verbs and adjectives as these grammatical categories are assumed to carry stronger lexical meaning than adverbs or prepositions. They should therefore be easily recognised when encountered in isolation.

b) Cognates, such as *detective* and *problem* (Uzun & Salihoglu, 2009), which are orthographically and/or phonologically similar to their translations in the learners' L1, were not included in any test material. One reason for not including cognates in the test materials is that the participants would have very little difficulty recognising these words due to the overlap between the target language and the learners' own language. Another reason is the likelihood of their eliciting an uncertain response behaviour in the testees, and therefore, leading them to overestimate their knowledge of vocabulary, although it was shown by previous research that cognates do not have a negative effect on the validity of a test if their number is close to the proportion that actually occur in the language (Meara et al., 1994).

In the tests employed, the target words were presented either in isolation or in bold within a context sentence which does not give hints to the target word meaning. Also, gender-biased language was avoided, or balanced gender representation was ensured.

3.3.2. The tests. In the present study, three different bilingual (English-Turkish) receptive vocabulary tests were used. These redesigned tests were based on the Vocabulary Size Test (Nation & Beglar, 2007a, 2007b), Vocabulary Levels Test (Nation, 2001) and Yes-No Test (Meara, 1992) in terms of format and most tested items. They each had a different format. For example, the VST was in a multiple choice format, the VLT was in matching format, while the Y/N Test was in a checklist format. As mentioned before, all the three tests aimed at measuring the first 5000 word families and included 18 target items which represented each of the five 1000 word families and which were selected on the basis of frequency.

In addition, at the beginning of each test, there were test instructions which were provided to the participants in their native language. The test instructions also included an example question in order to encourage understanding of each test format.

3.3.2.1. The vocabulary size test. In this test, the size of the learners' receptive vocabulary knowledge was measured using an adapted bilingual (English-Turkish) test with the same design features as the Vocabulary Size Test (Nation & Beglar, 2007a, 2007b). However, the frequency bands were mixed throughout the test. As mentioned in the literature, it was thought that mixing all the frequency bands would guarantee a good spread of low frequency words throughout the test; therefore, it would maintain engagement rather than cause the participants to give up at the lower frequency bands (Nation, 2012b; Nguyen & Nation, 2011). The pattern appeared in the redesigned VST was as follows: The first 18 target items, for instance, included six target word items from the tested band (1k here), and three items were taken from each of the 2k, 3k, 4k and 5k bands, constituting the remaining 12 target words. The second 18 items again included six tested items but this time from the 2k band and three items from each of the remaining four bands, and so on.

Each target word of the total 18 from a different frequency band was presented to the test takers in a short, non-defining sentence context followed by four answer choices, one of which was the correct equivalent of the target item and thus examinees must select as the correct one, while the rest were distractors chosen from the same frequency level as the test item. The target item and the non-defining context sentence it was put in were in the target language, in this study English, and the answer choices were in the native language of the participants, and they were given as either single-word or phrase-length Turkish definitions. Also, each answer choice was picked carefully as it needed to be equally plausible for the participants in the non-defining context sentence. The context sentence, on the other hand, would not assist the selection of the right answer written for each test item. In other words, the contexts were reflecting the most common environments for the test items. Two example item clusters are shown in Figure 8. As can be understood from the first item, each option seemed possible as it was an uncountable noun which grammatically fitted into the context sentence and collocated with the verb “use”. In the second item, all options shared the semantic feature of the word “feeling” and this time collocated with the verb “was”. In this case, test takers had to precisely know what the words “pressure” and “competent” referred to in order to answer the given target items correctly. You can also see an example item cluster from the online VST format in Figure 9 below.

pressure: They used too much <pressure>.

a. pastırma

b. tereyağı

c. basınç

d. nakit para

competent: She was <competent>.

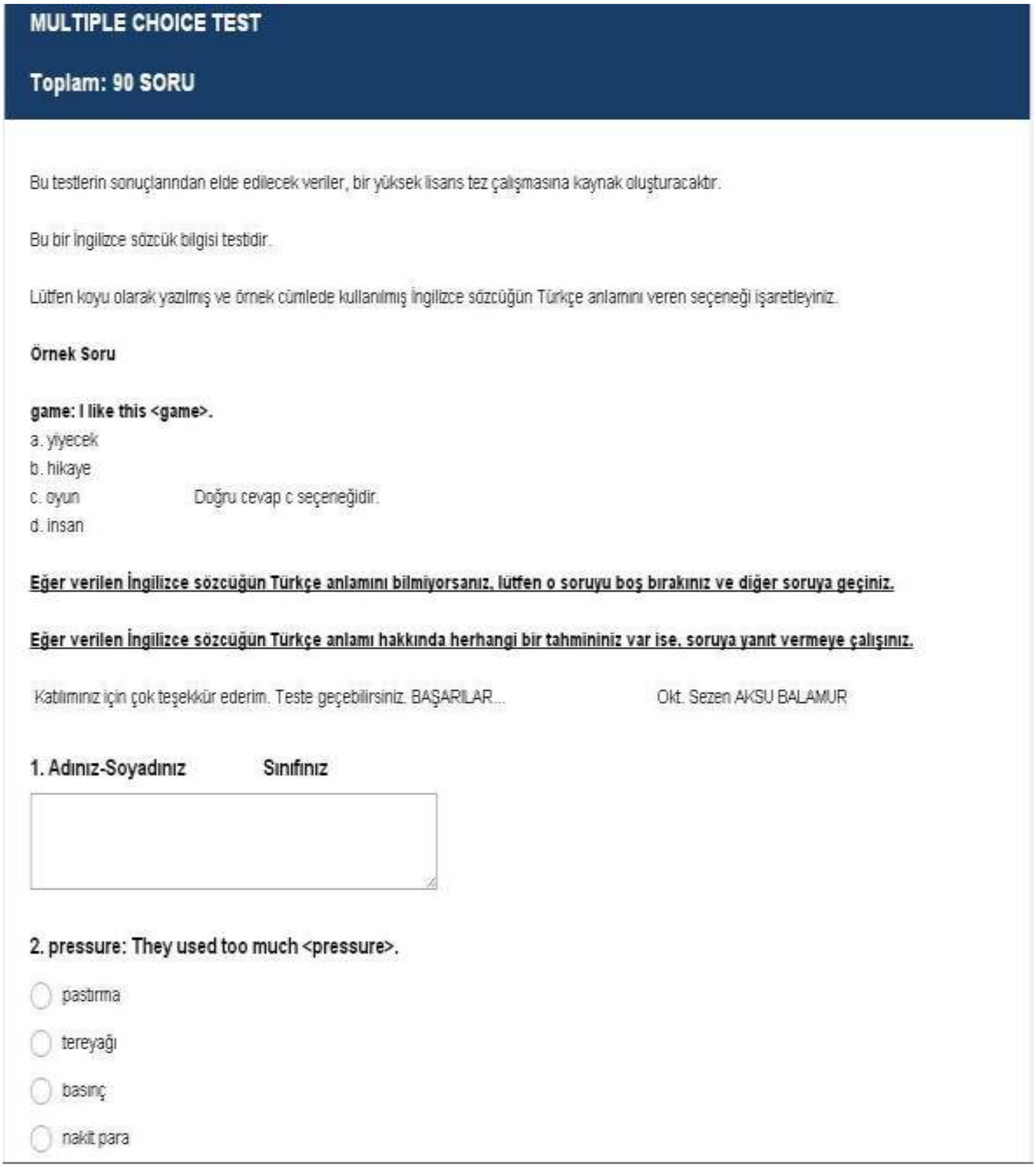
a. sadık

b. istekli

c. yetkin

d. savunmasız

Figure 8 Two example item clusters from the pen and pencil format VST.



MULTIPLE CHOICE TEST

Toplam: 90 SORU

Bu testlerin sonuçlarından elde edilecek veriler, bir yüksek lisans tez çalışmasına kaynak oluşturacaktır.

Bu bir İngilizce sözcük bilgisi testidir.

Lütfen koyu olarak yazılmış ve örnek cümlede kullanılmış İngilizce sözcüğün Türkçe anlamını veren seçeneği işaretleyiniz.

Örnek Soru

game: I like this <game>.
a. yiyecek
b. hikaye
c. oyun Doğru cevap c seçeneğidir.
d. insan

Eğer verilen İngilizce sözcüğün Türkçe anlamını bilmiyorsanız, lütfen o soruyu boş bırakınız ve diğer soruya geçiniz.

Eğer verilen İngilizce sözcüğün Türkçe anlamı hakkında herhangi bir tahmininiz var ise, soruya yanıt vermeye çalışınız.

Katılımınız için çok teşekkür ederim. Teste geçebilirsiniz. BAŞARILAR... Okt. Sezen AKSU BALAMUR

1. Adınız-Soyadınız Sınıfınız

2. pressure: They used too much <pressure>.

○ pastırma
○ tereyağı
○ basınç
○ nakit para

Figure 9 An example item cluster from the online format VST.

The participants were instructed first to read the English target word given in bold and the context sentence and then to circle the option which shows the Turkish definition of the target item. In order to reduce guessing effect, they were told to leave the question unanswered if they did not know the meaning of the target word. However, they were also told that if they had some partial knowledge about the target word, and if they had a guess about its meaning, they could have a try to find the correct option. In other words, in the VST, the testees were well-instructed not to make any guesses on the items which they lacked knowledge of. On the other hand, the instructions explicitly stated that the students should certainly try to find the correct choice if they thought they knew the meaning of the tested word or even part of it.

When it comes to scoring, each item in the test was given 1 point. In other words, participants received 1 point for their every correct answer. For instance, the test taker who answered all the test items correctly got 90 points in total.

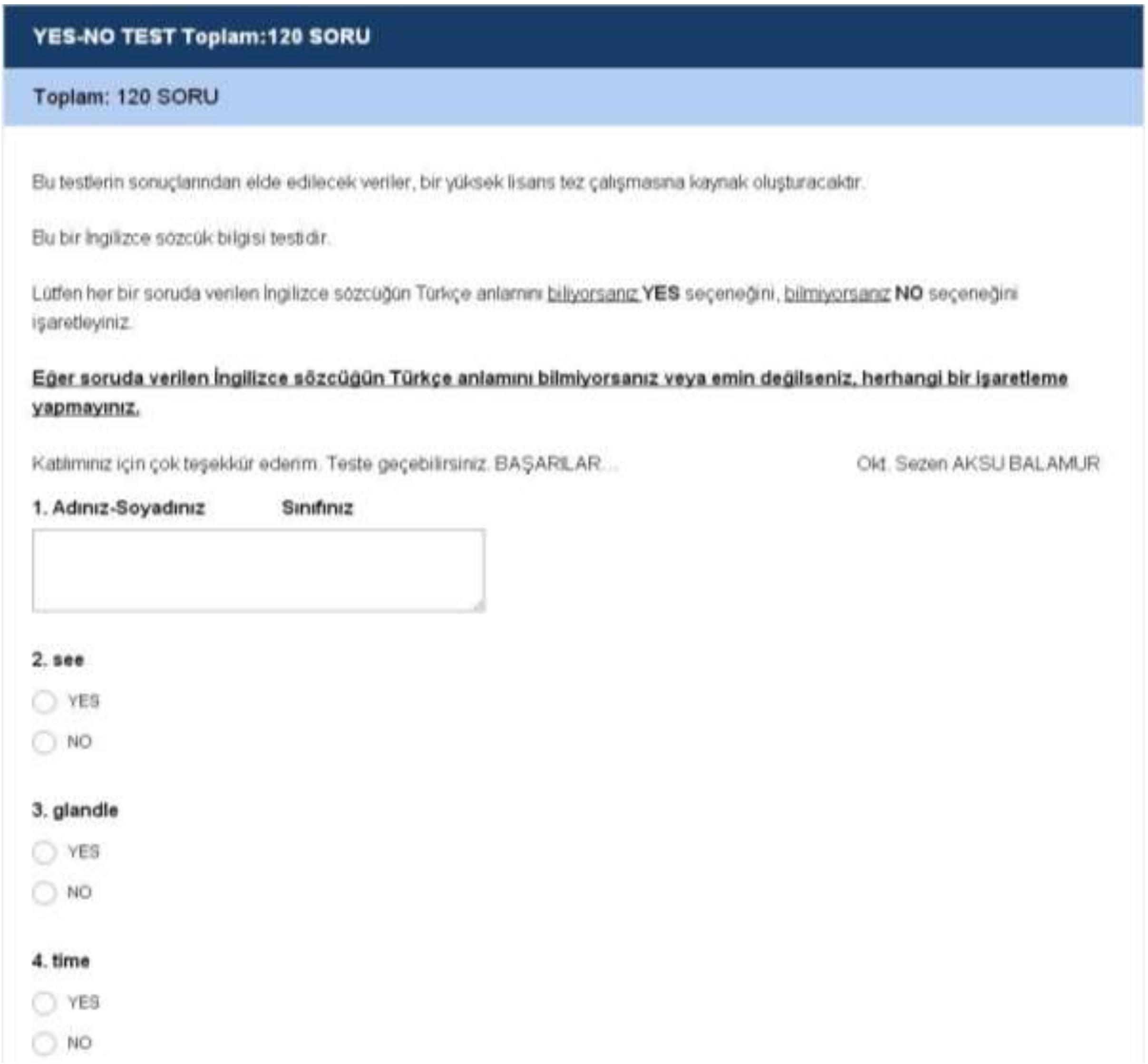
3.3.2.2. The yes-no test. The Yes/No test employed in this study was in the same fotmat as the one developed by Meara (1992); however, in order to make it paralel to the other two tests, it was constructed using the same 90 real-words used in the other two tests. In this test, along with 90 test items, there were 30 non-words in order to prevent overestimation. These non-words were taken from Meara's (1992) original Y/N Test following certain criteria. For instance, the ones which differ from real English words just because it had one different letter (e.g. "pring", which can be confused with "bring" and "rudge", which can be confused with "nudge") were not selected as they could have been too attractive to the participants to be rejected. The tested items were presented in three sets, each containg 30 target words and 10 non-words, which were randomly distributed. In each set, there were the same number of target items from each frequency level; that is, each set included six target words selected randomly from each of the first five frequency levels.

As for scoring the test items, like in the VST and VLT, the participants were awarded 1 point for each correct answer to the target item. As mentioned before, the same variable was used in Alderson's (2005) study, and similarly the non-words were ignored and the credited items were the real words which the participants identified correctly. Based on the figures reported by Alderson (2005), the raw hits can be regarded as an alternative indicator of overall vocabulary knowledge. On the other hand, there was a penalty for the examinees who checked more than three non-words in the Y/N Test. Such participants were eliminated from the study. From both institutions, in total there were 7 students from the elementary level, 10 students from the pre-intermediate level and 8 students from the intermediate level, who were excluded from the study.

As Figure 10 shows, in this test, examinees were instructed to put a tick in the box given next to the test item when they think they know its meaning. They were also told to refrain from marking the item as known if they do not know or they are not sure about its meaning. You can also see in Figure 11 below how example items were provided to the participants through the online VST format.

1□ see	11□ maintain	21□ crab	31□ review
2□ glandle	12□ litholect	22□ peasant	32□ strap
3□ time	13□ result	23□ acklon	33□ galpin
4□ threshold	14□ humberoid	24□ abandon	34□ latter
5□ connery	15□ poor	25□ vocabulary	35□ congest
6□ eclipse	16□ bodelate	26□ knee	36□ shoe
7□ deficit	17□ fragile	27□ dowrick	37□ compound
8□ adair	18□ batcock	28□ dig	38□ immigrate
9□ speech	19□ lonesome	29□ weep	39□ commemorate
10□scrub	20□ joke	30□ fracture	40□ seal

Figure 10 Example items from the pen and pencil format Y/N Test.



YES-NO TEST Toplam:120 SORU

Toplam: 120 SORU

Bu testlerin sonuçlarından elde edilecek veriler, bir yüksek lisans tez çalışmasına kaynak oluşturacaktır.

Bu bir İngilizce sözcük bilgisi testidir.

Lütfen her bir soruda verilen İngilizce sözcüğün Türkçe anlamını biliyorsanız **YES** seçeneğini, bilmiyorsanız **NO** seçeneğini işaretleyiniz.

Eğer soruda verilen İngilizce sözcüğün Türkçe anlamını bilmiyorsanız veya emin değilseniz, herhangi bir işaretleme yapmayınız.

Katılımınız için çok teşekkür ederim. Teste geçebilirsiniz. BAŞARILAR... Okt. Sezen AKSU BALAMUR

1. Adınız-Soyadınız **Sınıfınız**

2. see

○ YES

○ NO

3. glandle

○ YES

○ NO

4. time

○ YES

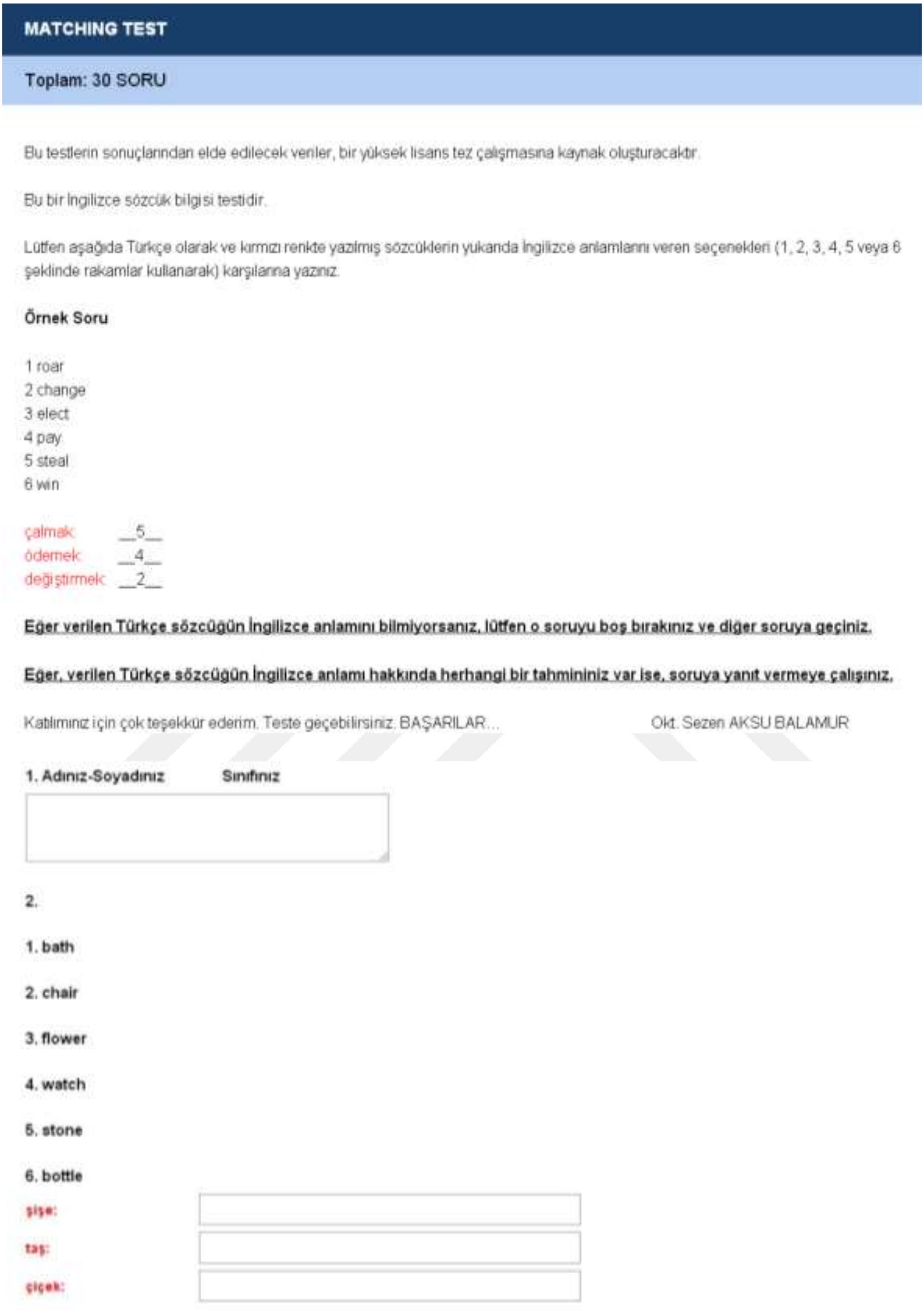
○ NO

Figure 11 Example items from the online format Y/N Test.

3.3.2.3. The vocabulary levels test. In the VLT, the participants were required to match three Turkish words with their English equivalents in each cluster containing six words. While organizing the test items, the three target words and the three distractors in each cluster were from the same fequency level and their part of speech was also the same. Also, the resercher was careful not to put tricky Turkish equivalents, such as *bottle* and *glass,* in the same cluster. The clusters were mixed thoroughout the test as was done in the other two tests. Target items in each cluster respresenting a different frequency band followed the order of 1k, 4k, 2k, 5k, 3k, 1k, 2k, 4k, 3k, 2k, 5k, 1k, 4k, 3k, 1k, 2k, 5k, 3k, 1k, 5k, 2k, 4k, 3k, 4k, 3k, 1k, 5k, 4k, 5k, and 2k. As in the VST, the first 18 target items (the first six clusters) in this test included six target word items (two clusters) from the tested band (1k here), and three items (one cluster) were taken from each of the 2k, 3k, 4k and 5k bands, constituting the remaining 12 target words (four clusters). The second 18 items covered two clusters from the 2k band and one cluster from each of the remaining four bands, and so on. Given below, Figure 12 shows an example item cluster used in the pen and pencil format VLT, while Figure 13 presents an example item cluster from the online VLT format.

1. appreciate
2. prove ___ **sürdürmek**
3. refuse ___ **uyarmak**
4. maintain ___ **reddetmek**
5. warn
6. select

Figure 12 An example item cluster from the pen and pencil format VLT.



MATCHING TEST

Toplam: 30 SORU

Bu testlerin sonuçlarından elde edilecek veriler, bir yüksek lisans tez çalışmasına kaynak oluşturacaktır.

Bu bir İngilizce sözcük bilgisi testidir.

Lütfen aşağıda Türkçe olarak ve kırmızı renkte yazılmış sözcüklerin yukarıda İngilizce anlamlarını veren seçenekleri (1, 2, 3, 4, 5 veya 6 şeklinde rakamlar kullanarak) karşılarına yazınız.

Örnek Soru

1 roar
2 change
3 elect
4 pay
5 steal
6 win

çalmak: __5__
ödemek: __4__
değiştirmek: __2__

Eğer verilen Türkçe sözcüğün İngilizce anlamını bilmiyorsanız, lütfen o soruyu boş bırakınız ve diğer soruya geçiniz.

Eğer, verilen Türkçe sözcüğün İngilizce anlamı hakkında herhangi bir tahmininiz var ise, soruya yanıt vermeye çalışınız.

Katılımınız için çok teşekkür ederim. Teste geçebilirsiniz. BAŞARILAR... Okt. Sezen AKSU BALAMUR

1. Adınız-Soyadınız Sınıfınız

2.

1. bath

2. chair

3. flower

4. watch

5. stone

6. bottle

şişe:
taş:
çiçek:

Figure 13 An example item cluster from the online format VLT.

In this test, the participants were instructed first to read the Turkish word given in bold and second to circle the option which shows the English form of it. As was done in the VST, they were also told to skip the question, but this time, when they did not know the English target that matches the Turkish definition. However, if they had some partial knowledge about the English target word and if they felt that they might know the correct answer, they were told to have a try.

Regarding scoring, the participants got 1 point for each correct answer to the test item. They did not have any penalties for their wrong answers.

3.3.3. The exit score. In this study, the exit score represents 50% of the formative in-term exam grades (progress tests students took during the year) including both terms and 50% of the proficiency exam grades of the participants. There were two quizzes and two midterms for each skill in a term. These exams include a wide range of question types such as cloze tests, multiple choice questions, fill in the blanks forms, matching questions, and open ended ones. The high-stakes paper-based proficiency exam takes place at the end of each academic year. It is made up of 80 multiple choice format questions as a major component measuring all the receptive and productive skills as well as the language elements, namely vocabulary and grammar, and a 20-point writing section, in which the students are required to write an opinion essay of about 250 words. All the institutional exams are prepared by the members of the Testing Office.

3.4. Data Collection Procedures

3.4.1. The pilot experiment. This study was done at the beginning of the second term at the university where the researcher works and was followed by a think-aloud procedure including 2 students: one from the elementary level group and the other from the intermediate level group. The aims of the pilot study and the think-aloud procedure were as follows: to realize how the students’ thinking processes proceed; to see how well target words and their

Turkish equivalents as well as distractors work; to spot any problematic or dysfunctional items; to highlight any unforeseen problems; and finally, to be able to make the necessary adjustments accordingly if necessary. It was also necessary to pilot the context sentence for each item to ensure that the test does not conflate the construct of the L2 contextual inferencing through vocabulary knowledge.

The pilot study was done in pen and pencil format in the classroom during the class time by the researcher herself in elementary level and one of her female colleagues in the pre-intermediate and intermediate levels. The instructions were given to the test takers in their native language (Turkish) orally. They were also presented as written at the beginning of each test.

In the pilot study, there were 48 examinees. 17 of them were elementary level students, the other 15 were from the pre-intermediate level, and the rest were from the intermediate level. Among these students, two volunteers were the ones who took the exams in a think-aloud format and who were recorded during the test sessions.

After the pilot study, one of the non-words*, "whitrow"*, had to be replaced with a new non-word, "*adair"*. It was the only non-word that was checked by the test participants more than the real target words, most probably because they consufed it with the real English word "withdraw" as a result of their limited lexical competence or limited access to a L2 lexicon.

3.4.2. Test Administration. The online testing software Survey Monkey <surveymonkey.com> was used to make it possible for the researcher to rapidly administer and analyze the tests with a large number of participants. The tests were administered either by the researcher herself or by her colleagues. Beforehand, the researcher informed her colleagues one by one about the test administration process through a short hands-on training. They were also given information on the test instructions although the instructions were all written in Turkish and presented at the beginning of each test.

The examinees took the tests within a week period in two sessions. In the first session, two tests were administered in one sitting (the first group took the Y/N Test & VST, and the second group took the Y/N Test & VLT, correspondingly), while in the second one, the participants continued with the remaining test (the first group took the VLT, and the second group took the VST), either in the room equipped with computers that the researcher arranged or in the classroom or garden (to reduce the participants' stress and to spark willingness in them) via cell phones.

First of all, before the test administration process started, all the students were informed about the aim of the study. Secondly, those who were volunteer to participate were asked to sign a consent form (see Appendix 4) to ensure that they would take part in this study willingly, and it would be possible for them to leave out whenever they wished. Following that, the participants were given as clear instructions as possible in their first language on what the correct procedure should be to prevent confusion. Then, they were divided into two groups, and the tests were given to them in a strict order within two sessions. One group was given the word-meaning recall based Y/N Test and the word-meaning recognition based multiple-choice format VST, while the other group took the Y/N Test and the word-form recognition based matching format VLT, correspondingly. The reason for administering the tests in different orders was to avoid or minimize any possible learning effect of the tested words from one test to the other as all the target items included in the Y/N Test also appeared in the other two tests. The participants were not given a time limit in order not to make the participants feel stressed and also to avoid any intervening variables; however, the first session took them about 20 minutes to answer the questions. In the second session, the participants who were given the Y/N Test and VST in the first session took the VLT, and those who took the Y/N Test and VLT were given the VST. Although there was again no time limit, the second session took almost 10 minutes. Thus, each participant did all three tests in two sessions and spending up to 30 minutes in total. Since the two test sessions were arranged to take place within a week, the second session was after three or four days following the first administration. This was the procedure followed at the first university where the researcher has been working. Yet, because of some institutional constraints, the participants at the second university had to do a pen and paper test in both sessions in their classrooms by following the same guidelines mentioned above.

3.5. Data Analysis Procedures

In the current study, quantitative data analysis procedures were utilized. After the data collection process was completed, most of the data which were obtained from the participants were automatically transferred into an Excel table from the online data collection application, surveymonkey.com, the main data collection tool for the study. Besides, manual data entering process was implemented for the rest of the data. The participants from three different proficiency levels were put together in the aforementioned three datasets, labeled differently as Elementary, Pre-intermediate, and Intermediate groups in accordance with their proficiency level. Once all the scores of 581 participants were obtained, analyses were conducted using the Statistical Package for the Social Sciences (SPSS) version 22.0. Then the quantitative data analysis procedures were followed to seek answers to the research questions.

With regard to the first three research questions whose aims were firstly to reveal whether the three vocabulary tests' results show any difference across proficiency levels; secondly to compare the three tests' results across test sections; and lastly to compare the three tests' results across different frequency and proficiency levels, the analysis of variance test (One-way ANOVA for repeated measures) was run. However, whenever an assumption of ANOVA was violated, two non-parametric tests, the Friedman Test and the Wilcoxon Signed Rank Test (for pairwise comparisons) were conducted.In addition, the Benforroni test was adopted for multiple comparisons. The aim of the last research question, though, was twofold: first, to correlate the three tests' results with one another to see which ones have the highest correlation level and second, to correlate the results of each test with the participants' university preparatory class exit scores. In order to answer this question the Spearman's Rho Calculator was run as the data was not normally distributed and the assumption of normality was not fulfilled.

This chapter presented as detailed information as possible on the aim of the study, participants, setting, data collection materials and steps as well as the procedures regarding the analysis of the data gathered. In the next chapter, the findings will be presented, and a comprehensive discussion of these findings will be included.

Chapter IV

Results and Discussion

In this part, firstly, the results obtained from a variety of statistical analyses of the quantitative data are presented. After the demonstration of the results under the guidance of each research question respectively, the findings are interpreted and discussed in detail with reference to previous research evidence.

4.1. Comparison of tests over proficiency levels

In the current study, the first research question was as follows:

"Do the three English vocabulary tests, the Vocabulary Size Test, Vocabulary Levels Test, and Yes-No Test, reveal similar overall receptive vocabulary size estimates in different proficiency levels?"

This research question required a comparison of the three test scores obtained from different proficiency groups. The scores of the tests are presented in Table 8. As seen from the table, in order to examine the differences between the test scores over proficiency levels, first, the descriptive statistics of the three tests' scores were calculated. The total score a participant could get was 90, and it was the same for all the three tests given as that many target words were tested in each test.

Next, the assumptions of the ANOVA were checked, and it was seen that the test scores for the elementary group Y/N Test and pre-intermediate group VST and VLT were not normally distributed ($p < .05$), as assessed by the Kolmogorov-Smirnov Test. What is more, the Mauchly's Test of Sphericity indicated that except for the elementary group, with the value ($p < .05$), the variances of differences between the test scores were significantly different. In other words the assumption of sphericity had been violated. On the other hand, as mentioned before, elementary group Y/N Test scores did not fit normal distribution in the Kolmogorov-Smirnov Test. As a result, instead of doing a repeated-measures ANOVA (as the assumptions were not met), the so-called (non parametric) Friedman Test was conducted for all the three proficiency groups. It was obvious from the results that there was a significant difference among the three test scores for all proficiency level groups (Elementary: [$\chi^2_{(2, N=211)}=47,44$, $p<0,01$]; Pre-Intermediate: [$\chi^2_{(2, N=259)}=110,73$, $p<0,01$]; and Intermediate: [$\chi^2_{(2, N=110)}=12,79$, $p<0,01$]). The results presented in Table 8 also showed that when the three test scores were compared across proficiency groups, the lowest scores were spotted in the elementary level. The scores of the pre-intermediate group were higher than the elementary group, and the intermediate group got the highest scores in all the tests they took.

Table 8

Comparison of Tests across Proficiency Levels

		VLT	VST	Y/N	Statistical Significance
Elementary	$\bar{x}$	38,568	38,658	34,388*	$\chi^2_{(2, N=211)}=47,44$ *
	SD	11,587	11,845	9,765	
Pre-Intermediate	$\bar{x}$	46,32	48,474	40,996	$\chi^2_{(2, N=259)}=110,73$ *
	SD	10,226	12,034	10,224	
Intermediate	$\bar{x}$	55,172	57,027	51,445	$\chi^2_{(2, N=110)}=12,79$ *
	SD	10,986	12,068	10,460	

* $p<0,01$

It was, as well, elicited from the results that differences at least between two proficiency groups existed. Therefore, in order to examine closely between which groups there was a significant difference, pairwise comparisons of the tests across proficiency groups were made using another non-parametric test, the Wilcoxon Signed Rank Test. The results provided by this test assured that the Y/N Test scores were significantly lower than the VST and VLT scores for all the three proficiency groups. This means that the Y/N Test could not precisely reflect actual lexical competence of the participants as it provided the lowest mean scores at all the proficiency levels. In fact, there may be some reasons that lie behind.

One of the reasons might be that though the Y/N task has a decontextualised format, the VLT and VST are cued recognition tasks, so they provide a partial context for responses. This may have caused the participants to perform better in the VLT and VST than they did in the Y/N test. As argued by Mochida and Harrington (2006), a miss or a “no” response to a target word will either accurately reflect the absence of any word knowledge or will underestimate what an individual actually knows and would be able to demonstrate when given more contextual support. Therefore, since the Y/N measure rely solely on “self-assessment”, the students’ actual knowledge of the target item cannot be verified, and as stated by some researchers, students’ self-estimates may sometimes be a poor indicator of the actual vocabulary knowledge they have (see Stubbe et al., 2010; Stubbe et al., 2011). Mochida and Harrington also argue that a central issue in vocabulary testing is what constitutes the word knowledge. With regard to the authors’ point of view, if the Y/N task is considered to be strictly a decontextualized vocabulary recognition measure, then we can say that there is not underestimation in the present study as three discrete states of knowledge were measured through the tests employed: words recognizable without any context; words recognizable merely with partial context, and words not recognizable in any way. Yet, the explicit supposition in the field is that knowledge of vocabulary is graded, and the Y/N format reflects that knowledge (Huibregtse et al., 2002).

The current study also reinforces the assumption that one reason for the low scores on the Y/N Test might be the students’ partial knowledge of some words. For example, during the think aloud protocol, the student from the intermediate group recognized the word ‘crab’. He said “I think this is a kind of animal which lives in the sea, but I am not sure; I may have confused it with another word”, and because of not being sure about the exact meaning of the word item, he rejected it. However, when he encountered the same item in the other tasks with an L1 translation, he did not hesitate to check it correctly. The same might have been the case for the other participants, which also means not only low-level students but also high-level ones may sometimes lack enough confidence in their lexical knowledge. In other words, the Y/N Test seemed to show signs of a blatant uncertainty within the test-takers when it came to make a decision on whether they actually knew the meaning of a tested word or not. This example further shows that the decontextualized items in the Y/N test, together with the pressure the instructions about future re-testing have put on them, seem to have the potential to prevent participants from successful guesses though they are better guided by their partial knowledge while answering the VLT and VST items. The above-mentioned student actually had enough partial knowledge to mark the item as known, but there were not any definitions like the ones in the other two tests to back up his partial recall. Related with this explanation, it might be better to consider the Y/N Test as a meaning-recall task as offered by Schmitt (2010), even though examinees do not have to demonstrate the target word meaning in the test. If so, then it is not unusual for a word meaning-recall task to be more difficult than a word meaning-recognition task required in the VST and a word form-recognition task required in the VLT as suggested by Laufer et al. (2004) and Schmitt (2010). This finding of the present study also seems to back up the assumption that “recall tests tend to underestimate vocabulary knowledge, ...” (Coxhead et al., 2014).

Another reason for rejecting a word item in the Y/N Test and answering it correctly afterwards can be the fact that the bilingual format of the VLT and VST tasks might have made them easier tests for the participants. This could partially account for the great amount of underestimation found in the direct comparison of the test scores presented in this study. The scores obtained in the current study are in accordance with ample evidence from previous research revealing that examinees score higher when given bilingual vocabulary tests (Elgort, 2013; Karami, 2012; Nguyen & Nation, 2011; Ruegg, 2007). The scores also back up Elgort’s argument that bilingual tests may reduce examinees’ anxiety, and therefore they might provide a “more accurate estimation of the breadth of their L2 vocabulary knowledge” (2013, p. 269).

Studies also show that the low level Yes/No test performance can be related with some other general factors such as the language background of the examinee, e.g. Dutch learners of French in Beeckmans et al. (2001) and Eyckmans (2004) had a lot higher FA rates than the Asian background ESL learners studied by Mochida and Harrington (2006); the number of target items as well as proportion of target words to pseudowords used (Beeckmans et al., 2001); the learner’s language proficiency level (Meara, 1996a), which is of particular importance, and the participant’s experience of the language skill being assessed, which also indicates that learners make use of recollections of their general language proficiency while making judgements (Ross, 1998). In addition, since subjects can respond so differently when they are faced with a Yes/No Test, even national characteristics may tend to emerge. For example, Shillaw (1999) reports that the Japanese learners in his study were very conservative in their own word knowledge estimates, so the nonwords were so rarely checked. It was reported by Al-Hazemi (1993) as well that Saudi students tend to use huge amounts of guessing strategy in this test type, possibly because their training was based on examination technique; nonetheless, the tests seemed to work reliably in these contexts. Eyckmans et al., (2007), studying with Belgian learners, for instance, reported a vast amount of overestimation (almost 60% of the nonwords were identified by the participants as real). The subjects of this study were required to finish the Yes/No Vocabulary Test. Afterwards, they were presented with the 60 existing Dutch words of this test and were asked to translate each test item in their own language. The results presented that nearly half of the real words evoking a Yes response in the preceding Yes/No Test were unfortunately translated incorrectly (Eyckmans et al., 2007:74).

In fact, such studies highlight another important factor in Turkish EFL classes. Though it is postulated by Meara (1996) that most learners acquire L2 words from exposure to the language, some cannot break the habit of trying to memorize lists of isolated words due to the fact that teachers also tend to focus on individual words as argued by Henriksen (2013) and Schmitt (2010). Using such a method limits the gain of broader knowledge about target words, which is reflected in lower Y/N Test scores in the study. In order to avoid this, teachers had better use vocabulary activities which trigger contextual learning.

Based on all these, it can be said that various factors might have influenced the performance of the subjects included in this study and cause them to underestimate their knowledge of vocabulary in the Yes/No Test. On the other hand, whereas acknowledging the role such factors could play, the Yes/No format is considered by researchers like Read (2000) to be a valid and potentially useful tool to estimate learners' L2 vocabulary knowledge.

When it comes to the pairwise comparisons of the VST and VLT scores across proficiency groups, even though the scores of the both tests were close to each other, the participants' VST scores were higher than their VLT scores at all proficiency levels. It was also obvious that at pre-intermediate proficiency level, though being small, there was a significant difference between the VST and VLT scores as given in Table 9.

Table 9

Pairwise Comparisons of Tests across Proficiency Groups

	VST-VLT	VST-Y/N	VLT-Y/N

Elementary	z=-0,288 n.s.
	VST=VLT
		z=-6,086 *
		VST>Y/N
			z=-6,428 *
			VLT>Y/N

Pre-Intermediate	z=-2,760 * VST>VLT	z=-8,998 * VST>Y/N	z=-7,560 * VLT>Y/N

Intermediate	z=-1,692 n.s.
	VST=VLT
		z=-4,189 *
		VST>Y/N
			z=-3,766 *
			VLT>Y/N

* p<0,01

If we remember Laufer et al.'s categorization of vocabulary knowledge once again, then it was not unusual to get the lowest scores from the Yes/No Test as in Schmitt's point of view, it should be considered as a meaning recall task (passive recall). Following Laufer's categorization, next comes the VLT, which is considered by Schmitt (2010) to be a form recognition task (active recognition) and lastly the VST, which is accepted by Schmitt as a meaning recognition task (passive recognition) and regarded as the basic level of vocabulary knowledge.

Related with this explanation, the reason for the VLT's offering lower scores than the VST for the pre-intermediate level participants may be that this level is a kind of bridge between elementary and intermediate levels, so the students' knowledge of vocabulary is neither as limited as the learners from the elemantary level nor as good as the ones from the intermediate level. That is why they could have felt more secure with the VST, and therefore got higher scores since it presents them limited items in the target language compared to the VLT. In the VLT, the test instructions made the students focus first on the meaning of the target words and then find the target word form. This means they were supposed to handle six different target words and recognise the form of each in order to mark the correct word form matches after reading the definitions of just three of them. However, in the VST, the students had just one target word form to deal with, and this was possibly an advantage the test offered to the examinees.

In addition, as argued by Laufer et al. (2004) and Schmitt (2010), going from word meaning to word form (form recognition/active recognition), a task required by the VLT, is a more difficult skill than going from word form to word meaning (meaning recognition/passive recognition), which is what the VST required from the test takers in the pressent study. As a result, such differences in the given test tasks might be the reason for the significant difference in the test scores of the pre-intermediate group, which meant the VST was slightly an easier measure for them than the VLT and also supported the ‘‘strength of vocabulary knowledge’’ hypothesis offered by Laufer and Goldstein (2004). A more detailed discussion regarding the comparison of the VLT and VST scores will be provided in the following sections of this thesis.

4.2. Comparison of tests across test sections

The second research question of the study was as follows:

“Do the VST, VLT and Yes/No Test reveal similar overall receptive vocabulary size estimates across test sections?”

This research question aimed at a comparison of the three tests’ scores in different test sections. The first step of the data analysis process to seek an answer to the second research question was again calculating the mean values for each test in each test section through descriptive statistics. The results of the tests over the five test sections are given in Table 10.

Following this step, the assumptions of ANOVA were tested in all five test sections (1K-5K) for the whole proficiency groups. Tests of Normality were used to examine whether there was a normal distribution among the test scores between 1K-5K test sections. The Kolmogorov-Smirnov Test results showed that the scores were not normally distributed in any test sections. Next, it was necessary to determine whether there was a significant relationship between at least two of the variances of differences for each test. Therefore, the Mauchly's Test of Sphericity was run. This test showed that when test scores were compared across 1K-5K test sections, with the value ($p < .05$), the results were statistically significant, which meant the assumption of sphericity was not met except for 3K section ($p > .05$). For all those reasons, while comparing the tests across test sections for 1K-5K sections, except for 3K section, the Frieadman Test was run instead of the repeated-measures ANOVA. In order to compare the test scores in 3K section, even though the Kolmogorov-Smirnov Test results did not seem to exhibit normal distribution, it was possible to run one-way repeated-measures

ANOVA as the values of the mean, median and mode were close, the values for skewness and kurtosis were less than ± 1.0, and most importantly, the condition of sphericity was met (*p* > .05), so pairwise comparisons were conducted as well.

According to the statistical results, when all the test scores were compared, there appeared to be a significant difference between at least two of them in all test sections: 1K [$\chi^2_{(2,\ N=580)}$=300,029, p<0,01]; 2K [$\chi^2_{(2,\ N=580)}$=109,226, p<0,01]; 3K [$F_{(2-1158)}$=10,678, p<0,01]; 4K [$\chi^2_{(2,\ N=580)}$=169,928, p<0,01]; and 5K [$\chi^2_{(2,\ N=580)}$=166,468, p<0,01].

Table 10

Comparison of Tests across Test Sections

		VLT	VST	Y/N	Statistical Significance
1K	$\bar{x}$	16,72	15,97	15,41	
	SD	2,102	2,045	2,128	$\chi^2_{(2,\ N=580)}$=300,029 *
2K	$\bar{x}$	11,25	11,33	10,07	
	SD	3,562	3,646	3,646	$\chi^2_{(2,\ N=580)}$=109,226 *
3K	$\bar{x}$	6,87	7,36	6,88	
	SD	3,299	3,223	2,640	$F_{(2-1158)}$=10,678 *
4K	$\bar{x}$	6,13	6,57	4,88	
	SD	3,074	3,502	2,950	$\chi^2_{(2,\ N=580)}$=169,928 *
5K	$\bar{x}$	4,21	5,30	3,34	
	SD	2,778	3,362	2,265	$\chi^2_{(2,\ N=580)}$=166,468 *

* p<0,01

Lastly, in order to ascertain between which tests a significant difference would be traced, the Wilcoxon Signed Rank Test was conducted in 1K, 2K, 4K, and 5K sections. In 3K section, though, Bonferroni pairwise comparisons were performed.

The statistical results (as given in Table 11) displayed that when the Y/N Test scores were compared to VST and VLT scores across test sections, Y/N Test scores were significantly lower than both VST and VLT scores at all frequency levels except for 3K section, where VLT and Y/N scores were close to each other and therefore lacked a significant difference. On the other hand, VST-VLT comparison made it clear that while VST scores tended to be lower than VLT scores in higher frequency test sections (i.e.1K) as the participants proceeded to lower frequency test sections (3K, 4K, and 5K frequency bands), their VST scores got significantly higher than their VLT scores. These differences reached statistical significance in all test sections except 2K, which can be seen as the transition point for this change.

Table 11

Pairwise Comparisons of Tests across Test Sections

	VST-VLT	VST-Y/N	VLT-Y/N
1K	z=-10,520 * VST<VLT	z=-7,301 * VST>Y/N	z=-13,920 * VLT>YN

2K	z=-0,535 n.s.
	VST=VLT
		z=-9,267 *
		VST>Y/N
			z=-8,855 *
			VLT>YN

3K	VST>VLT *	VST>Y/N *	VLT=Y/N n.s.
4K	z=-3,366 * VST>VLT	z=-12,070 * VST>Y/N	z=-10,600 * VLT>YN
5K	z=-8,481 * VST>VLT	z=-13,012 * VST>Y/N	z=-7,334 * VLT>YN

* p<0,01

The detailed comparison of the tests across test sections (see Table 10) again showed that the Y/N Test scores were significantly lower than the other two test scores in almost all test sections. This finding brings along the idea that it is highly possible for the participants to create susceptibility to warnings in the Yes/No Test instruction, and this susceptibility is evidently independent of the participants' vocabulary knowledge. The future re-test warning might have even scared the participants into becoming too conservative. There are a number of previous studies showing that the nature of the test instructions might influence some students and cause them to become too conservative in their knowledge of vocabulary by not checking the target words they actually know (underestimation) (e.g. Eyckmans, 2004; Eyckmans et al., 2007; Mochida & Harrington, 2006), which was also the case in the current study. For example, in a study by Abels (1994, as cited in Eyckmans, 2004), whose design was different from the present one, the testees were given the same Yes/No test twice. In the first experiment, the participants were not mentioned about the inclusion of nonwords in the test, but in the second one, they were told about the presence of nonwords in the word lists. Additionally, the participants were said that it was possible for them to alter their responses given in the first experiment from a "Yes" to a "No" response. According to the result of the study, it was reported that once the participants were informed about the nonwords, they did not tend to overestimate their vocabulary knowledge. Furthermore, there was a decline in the number of "Yes" responses not only in the nonword but also in the real word category though more changes were spotted in the nonword category than in the actual word category. This finding shows that the test instructions had clearly caused the participants to chose a more careful response behaviour, which was also the probable reason for the low Yes/No scores obtained from the present study since the researcher might have intervened in the participant's decision making process by reinforcing the test's instruction about future re-testing.

In another study, Mochida and Harrington (2006) assessed their advanced level subjects' Yes/No test performance as a predictor of their subsequent performance on the VLT. Including the same VLT items in the Yes/No formats, the researchers were able to check directly for instances of overestimation and underestimation of the students' word knowledge in the Yes/No tests. The authors determined that contrary to expectations, there were no traces of overestimates in the study. On the other hand, overall there was not a link between the low FA rates and underestimates on the VLT, except for the 5K-10K bands. This finding suggested that the subjects who were more conservative in giving a "Yes" response to nonwords did not have more underestimates, a result that does not match with the current study. According to the authors, the reason for the participants' underestimation of their vocabulary knowledge at lower frequency bands, though, might be the warning taking place in the test instructions. They warned their participants that they might be retested on their knowledge of some of the target words after they had taken the Yes/No test. This specific criterion, as the authors suggest, may have resulted in lower FA rates and closer scores to those of the VLT. The findings of the present study are thus in line with the above-mentioned work in that though the main purpose here was not to check either for FA rates or the FA correlations with the number of overestimates or underestimates, it was clear that the number of the students who had more than three FAs was rather small (25 students in total). In addition, significant differences between the Yes/No Test scores and these of the VST and VLT meant that the students were underestimating their knowledge of vocabulary in the Yes/No Test. Regarding this finding, it can be concluded that this study does not support Mochida and Harrington’s (2006) statement that “The results show that the Yes/No test is a reliable measure of the kind of vocabulary knowledge measured by the VLT and, presumably, similar multiple-choice tests” (p. 91). In this respect, this study also seems to concur with the concern felt by Sims (1929) about the Y/N test’s power in giving a satisfactory measure of vocabulary as the low Y/N test results in this study did not accurately reflect the testees’ word knowledge. However, the results are not in line with the study by Sims which suggests that Y/N test results overestimate the number of words actually known by test-takers. We suggest just the opposite - that Y/N test results underestimated the number of words known by the students. The findings possibly suggest exceedingly cautious performance providing scores that under represent the participants’ actual knowledge of vocabulary.The number of underestimates is considered by some researchers to be informative in that they provide evidence about whether learners who are more conservative and accordingly who have lower FA rates, are also more conservative in identifying real words (e.g. Huibregtse et al., 2002:231). This condition was in fact reflected in the current study as the participants’ scores were lower on the Yes/No Test regarding the real words and most of the participants did not exceed the non-words threshold which was limited to maximum three items.

This study further aligns with Mochida and Harrington’s in that similar instructions were provided to the participants involved. In the Yes/No test, the subjects of the present study were instructed to make a judgement on the basis of their knowledge of the basic word meaning regarding the fact that functional knowledge of a word can include “the ability to recognize the form, translate it into the L1, recognize it in context, or accurately use it in an informative context, among other things” (Mochida & Harrington, 2006). Like in Mochida and Harrington, the participants here were also warned that their knowledge of the same words would be checked through two different tests after they had done the Yes/No test, a condition that was missing from some earlier studies (Beeckmans et al., 2001; Cameron, 2002; Meara & Buxton, 1987) as it was thought that it might cause the participants to become too conservative and lead them to underestimate their word knowledge (see Stubbe, 2012). The low scores obtained from the test may actually have attributed to the participants’ variability in their judgment behavior as they were evidently prone to underestimating their knowledge of vocabulary in the Yes/No Test. On the other hand, it is assumed that the additional instructions about retesting the same words through different test formats subsequently may have caused the participants to maintain a more conservative approach and accordingly contributed to the underestimations of actual vocabulary knowledge evident here. Since response style is an individual trait, a subject with a conservative response style is not expected to say “yes” very quickly to either a pseudoword or real word.

Taking such a warning into account, this study is also similar to the one done by Eyckmans (2004) on lower intermediate level French-speaking university students (n=179) because in that study as well, the participants were warned that they would be re-tested on the Yes/No test items later. In Eyckmans’s study, using identical test material, the author investigated the relation between two different test instruction conditions and how they influenced the test-takers’ response behavior. In the minimal instruction condition participants were only asked to indicate whether or not they knew the word, while in the strict condition they were also announced that they would be tested on the Yes/No items later to check for the validity of their responses. With the latter instruction, the participants were urged to avoid ticking a tested word unless they were completely sure of their responses. Then the participants were asked to translate the French words that were previously presented in the Yes/No format into their own language. It was established that the FA rate diminished significantly when participants were provided a strict instruction, which means the response behavior of the participants was particularly influenced regarding the nonwords and yet not as much regarding the real words. In two previous studies carried by Eyckmans, though, there were high FA rates which meant the examinees were overestimating their knowledge of vocabulary. In the first experiment, the task of the students was to decide if they had ever encountered the words given, whereas in the second one their task was to determine if they knew the meaning of these words. In the latter task, the participants were asked to give the French translation of the existing words presented to them in the Yes/No Test. However, they were not informed beforehand about this second test to prevent a possible taint on their Yes/No Test performance. The comparison of the results of those both experiments concluded that the modified instructions neither reduced the FA rate nor inceased the validity or reliability of the test. These two experiments also revealed that high false alarm rates were not restricted only to weaker (beginner level) participants, but rather were often displayed by more proficient (advanced level) individuals. That is, even advanced level students who were better at identifying actual words did not show a better performance at rejecting pseudowords. Another important pattern in the second experiment was the percentage of the word-items (almost 17%) that elicited a “No”-response in the Yes/No Test, but were offered a correct translation afterwards. The data analysis showed that one out of every four word-items (both cognates and non-cognates) fell into this pattern. In addition, there was a difference in the “No” response + correct translation pattern for the third experiment, in which the tendency to reject existing words yet translate them correctly afterwards was more than twice as large. In

Experiment 3, approximately 37% of the actual words were given a “No” response, but later again 25% of these rejected word-items resulted in a correct translation. This difference between the experiments creates the impression that in Experiment 3, the participants were often tending to underestimate their word knowledge rather than overestimate it as they responded with more caution.

When these three experiments are compared to one another, it is clear that alternative, or manipulated, test instructions can influence the participants’ response behavior and can cause the test to address a different level of the students’ vocabulary knowledge. However, according to Eyckmans, with manipulated variables, it does not seem possible to overcome or counterbalance the inherent problem of the Yes/No format, as the task aims to measure two dimensions at the same time; namely the participants’ overall vocabulary size and their own estimation of their vocabulary knowledge. Regarding Eyckmans’ studies mentioned above, the present study aligns with them in that test instructions including a warning of possible re-testing of the same test items may have caused the participants to adopt a more conservative approach. In the current study, the students obviously had a tendency to give a “No” response even to existing words in the Yes/No task, but afterwards they were able to give a correct answer for these rejected items in the subsequent VLT and VST. This response behavior with regard to underestimates of real word knowledge seems to have justified the conservative approach taken by the participants. A different instruction could have influenced the learners’ responses in a different way, especially in the Yes/No Test format, where the testees’ responses bear ambiguous status. With the present study, it was once again clear that the participants’ response behavior was based not only on their lexical knowledge but also on their individual decision making process. Therefore, this study highlighted the fact that the implications of the test instruction on the individual’s choices should be paid more attention while assessing their vocabulary knowledge. As stated by Beeckmans et al. (2001), even if it can be assumed that the Yes/No test is tapping into a kind of fundamental word knowledge, such an assumption cannot rule out the likelihood of complex interaction taking place between different test instructions and several levels of knowing a word. In fact, the question of how and to what extent varying test instructions affect the performance of participants deserves further attention and investigation because of the fact that the instructions given to students before the test can work in unexpected directions. They can control and/or decrease the rate of overestimation, or conversely increase it, or they can cause students to underestimate their word knowledge. Then, in agreement with Eyckmans (2004), it can be said that influencing examinees' response behaviour by different test instructions that urge them to a more thoughtful or careful response behaviour in the Yes/No format does not automatically make the test a more valid vocabulary size measure.

4.3. Comparison of tests across frequency and proficiency levels

The third research question of the study was as follows:

"Do the VST, VLT and Yes/No Test reveal similar overall receptive vocabulary size estimates in different frequency bands for different proficiency levels groups?"

The aim of the third research question was to compare the three tests' results across different frequency and proficiency levels. As seen from Table 12, where the results of the tests were given, in order to seek an answer to this question, first, the descriptive statistics were calculated, and then the tests of normality were done for each test and proficiency group through 1K-5K frequency sections separately. The results showed that in 1K level, for all proficiency groups; in 2K level, for elementary and pre-Intermediate groups; in 4K level, for intermediate group; in 5K level, again for all proficiency groups, it was seen that the normality conditions, as assessed by Kolmogorov-Smirnov Test, and sphericity assumptions, as assessed by the Mauchly's Test of Sphericity, were not met. Therefore, the Friedman Test instead of ANOVA and the Wilcoxon Signed Rank Test were done respectively. On the other hand, in 2K level, for intermediate group; in 3K level, for all proficiency groups; and in 4K level, for elementary and pre-intermediate groups, although the Kolmogorov-Smirnov Test results did not seem to exhibit normal distribution, it became possible to conduct one-way repeated-measures ANOVA since the values of the mean, median and mode were close, the values for skewness and kurtosis were less than ± 1.0, and most importantly, the sphericity assumption was met ($p > .05$), and then pairwise comparisons were conducted as well.

The statistical test results showed that when the test score averages of the three proficiency groups were compared through 1K-5K levels, at least two were proved to be significantly different from each other for all proficiency groups in all test sections excluding the intermediate group in 3K level. For that group, no significant difference among the test score averages emerged, and thus there was no further need for pairwise comparisons (see Table 12).

Table 12

Comparison of Tests across Frequency and Proficiency Levels

			VLT	VST	Y/N	Statistical Significance
1K	Elementary	x̄	15,83	14,91	14,34	
		SD	2,631	2,477	2,319	$\chi^2_{(2, N=211)}$=95,544 *
	Pre-Intermediate	x̄	17,1	16,39	15,67	
		SD	1,66	1,470	1,773	$\chi^2_{(2, N=259)}$=180,923 *
	Intermediate	x̄	17,55	17,03	16,84	
		SD	1,046	1,237	1,345	$\chi^2_{(2, N=110)}$=31,756 *
2K	Elementary	x̄	9,32	9,22	8,27	
		SD	3,393	3,354	3,198	$\chi^2_{(2, N=211)}$=29,481 *
	Pre-Intermediate	x̄	11,54	11,82	10,13	
		SD	3,025	3,256	3,342	$\chi^2_{(2, N=259)}$=65,723 *
	Intermediate	x̄	14,29	14,25	13,37	
		SD	2,610	2,451	2,678	$F_{(2-218)}$=7,357 *
3K	Elementary	x̄	5,36	5,8	5,8	
		SD	2,943	2,929	2,478	$F_{(2-516)}$=4,138 *
	Pre-Intermediate	x̄	7,13	7,75	6,83	
		SD	2,942	2,987	2,373	$F_{(2-516)}$=12,497 *
	Intermediate	x̄	9,15	9,42	9,06	
		SD	3,280	2,865	2,185	$F_{(2-218)}$=0,773 n.s.
4K	Elementary	x̄	4,59	4,52	3,21	
		SD	2,705	2,661	2,016	$F_{(2-420)}$=39,680 *
	Pre-Intermediate	x̄	6,48	7,22	5,17	
		SD	2,685	3,156	2,560	$F_{(2-516)}$=70,424 *
	Intermediate	x̄	8,24	8,95	7,37	
		SD	3,103	3,610	3,323	$\chi^2_{(2, N=110)}$=17,734 *
5K	Elementary	x̄	3,47	4,22	2,77	
		SD	2,371	2,777	1,814	$\chi^2_{(2, N=211)}$=49,508 *
	Pre-Intermediate	x̄	4,07	5,29	3,19	
		SD	2,577	3,183	1,996	$\chi^2_{(2, N=259)}$=92,279 *
	Intermediate	x̄	5,95	7,37	4,80	
		SD	3,214	3,821	2,942	$\chi^2_{(2, N=110)}$=26,982 *

* $p<0,01$ * $p<0,05$

According to the mean scores of the tests reported above, frequency was an invariably significant factor in receptive vocabulary knowledge since in all the tests used in this study, each proficiency group did better in higher frequency bands whereby more words were known and the scores showing the number of the words known decreased linearly as the participants proceeded towards lower frequency bands. This finding of an inverse relationship between knowledge of vocabulary and level of frequency backed up the common assumption that higher frequency words are easier and thus learned faster than lower frequency words. Therefore, the current study seemed to indicate a developmental order in the participants' word knowledge in terms of frequency, such as having grater mastery of the 3K level than the 4K level, although a claim that the participants had full mastery of any given frequency level was not made. In addition, it was not possible to define the 5K level in this way since the levels beyond that one were not tested. Regarding this finding, the present study aligned with some previous studies providing evidence for learning rate differences between frequency levels (Laufer et al., 2004; Laufer & Paribakht, 1998; Milton, 2009; Richards et al., 2008) and confirming results which present a linear decrease over word frequency bands either in monolingual size tests (David, 2008; Milton, 2009; Ozturk, 2013, 2015, 2016; Richards et al., 2008) or in some bilingual ones (Elgort, 2013; Karami, 2012; Nguyen & Nation, 2011; Zhao & Ji, 2016) .

This finding also showed that the learners' vocabulary knowledge rate varied not only with the frequency level of the words but also with the profiency level of the participants as can be seen in Table 12. According to the means for all the three groups of learners, it can be said that the effect of proficiency was very strong on vocabulary learning because the learners' scores decreased in accordance with their level of proficiency; that is, the scores of the intermediate group were higher than the scores of the pre-intermediate group, and the scores of the pre-intermediate group were higher than the elementary group in all the tests.

With reference to the finding that more proficient students knew more words than less proficient students in each frequency band, the current study aligns with the research evidence provided by previous studies suggesting that the number of target items known has a direct relationship with the learners' level of ability, decreasing along with their proficiency (Stubbe, 2012) and that vocabulary knowledge can be regarded as a reliable proxy for general foreign language proficiency (Elmasry, 2012; Gyllstad, 2007; Koda, 1996; Maskor & Baharudin, 2016; Milton, 2009, 2013; Milton et al., 2010; Stæhr, 2008). This study also seems to reflect the possibility that a vocabulary size test might be a useful tool to make assumptions about general language proficiency, as such tests are claimed to provide the same information as other language proficiency measures (Milton, 2009).

Finally, as seen from Table 13, the results of the Wilcoxon Signed Rank Test and pairwise comparisons' showed that when the test scores were compared by taking the proficiency groups into consideration, the results were as follows: When the VST and VLT scores were compared with the Y/N test scores, it was seen once again they were significantly higher than the Y/N scores for all the three groups in all the test sections, except for the elementary and pre-intermediate groups in the 3K section (VST=Y/N n.s. and VLT=Y/N n.s., respectively) and the intermediate group in the 1K section (VST=Y/N n.s.). On the other hand, when the VST and VLT scores were compared, it was again obvious that in higher frequency levels, the VST scores were significantly lower than the VLT scores, yet in lower frequency levels, they got significantly higher than the VLT scores for both the pre-intermediate and intermediate groups. The VST-VLT comparison for the elementary group, though, showed that the VST scores were significantly lower than the VLT scores in higher frequency levels (1K; 2K though in 2K section, the difference did not reach a significant value); and as they went on towards the lower frequency levels (3K), the VST scores exceeded the VLT scores at a significant value, so they seemed to follow the same pattern as the other two groups.

However, in the 4K test section, they fell out of that pattern as the VST scores were once again lower than the VLT scores though the difference did not reach a significant value.

Table 13

Pairwise Comparisons of Tests across Frequency and Proficiency Levels

		VST-VLT	VST-Y/N	VLT-Y/N
1K	Elementary	z=-6,096 *	z=-3,938 *	z=-8,375 *
		VST<VLT	VST>Y/N	VLT>Y/N
	Pre-Intermediate	z=-7,587 *	z=-6,435 *	z=-10,033 *
		VST<VLT	VST>Y/N	VLT>Y/N
	Intermediate	z=-4,362 *	z=-1,397 n.s.	z=-4,822 *
		VST<VLT	VST=Y/N	VLT>Y/N
2K	Elementary	z=-0,261 n.s.	z=-4,608 *	z=-5,151 *
		VST=VLT	VST>Y/N	VLT>Y/N
	Pre-Intermediate	z=-1,259 n.s.	z=-7,464 *	z=-6,265 *
		VST=VLT	VST>Y/N	VLT>Y/N
	Intermediate	VST=VLT n.s.	VST>Y/N *	VLT>Y/N *
3K	Elementary	VST>VLT *	VST=Y/N n.s.	VLT<Y/N *
	Pre-Intermediate	VST>VLT *	VST>Y/N *	VLT=Y/N n.s.
	Intermediate	(no need for pairwise/ multiple comparisons)	(no need for pairwise/ multiple comparisons)	(no need for pairwise/ multiple comparisons)
4K	Elementary	VST=VLT n.s.	VST>Y/N *	VLT>Y/N *
	Pre-Intermediate	VST>VLT *	VST>Y/N *	VLT>Y/N *
	Intermediate	z=-2,126 *	z=-4,028 *	z=-3,103 *
		VST>VLT	VST>Y/N	VLT>Y/N
5K	Elementary	z=-4,239 *	z=-7,727 *	z=-4,125 *
		VST>VLT	VST>Y/N	VLT>Y/N
	Pre-Intermediate	z=-6,122 *	z=-8,987 *	z=-5,040 *
		VST>VLT	VST>Y/N	VLT>Y/N
	Intermediate	z=-4,049 *	z=-5,671 *	z=-3,450 *
		VST>VLT	VST>Y/N	VLT>Y/N

* p<0,01 * p<0,05

Turning back to the detailed comparison of the VLT and VST scores, as seen in Table 13, in general the VST seemed to be a more difficult tool for participants compared to the VLT at the higher frequency levels. However, it got easier for them at the lower levels. On the other hand, the opposite can be said about the VLT. The test appeared to be easier at the higher frequency bands, while at the lower bands, it got difficult probably because it became for students progressively harder to provide correct answers with decreasing frequency. Therefore, it can be said that as the frequency level based comparisons of these tests show, there is an inverse relationship pattern between them, which means the VLT and VST are not equivalent. This finding might be related to the reasons discussed below.

The first reason might be the assumption that the VLT has a more difficult and demanding format than the VST as the test requires students to read the meanings first and then look for the correct word forms given to them. That is why it can be considered here as a form recognition task as Schmitt suggests, which requires a higher degree of vocabulary knowledge. Also, in their study, Laufer et al. (2004), defines this type of word knowledge as active recognition and according to their participants' MC test results, this level knowledge is more difficult than passive recognition, the knowledge type that the VST in the present study aimed to measure.

The second reason might be that each cluster in the VLT contained target words and distractors which belonged to the same word class, and both the three tested words and the three distractors were from the same frequency band. This means participants had to deal with six words in total from the particular band in the given cluster. The reason for the participants' higher scores at higher frequency bands might be that the L2 words given in each cluster were items whose meanings were very different from one another. Consequently, the students with even a minimal idea of each target word meaning might have been able to match the given L1 meaning with the correct L2 word form. However, the scores got lower as participants proceeded towards the lower frequency bands. It was obvious that as the frequency band got lower, word item difficulty increased and this resulted in lower scores in the VLT because of the participants' lack of even partial knowledge about the given low frequency target words. Compared to the VLT, in VST, the participants had to focus on just one word form. They were instructed to read the non-defining example sentence including the target word first and then choose the correct meaning of the tested word from the four options. As it was mentioned before, this type of knowledge was regarded as the most basic level of vocabulary knowledge within their categorization by Laufer et al. (2004). The reason for lower scores at higher frequency bands in VST though might be that since in each test item all the options presented to the participants were defining some other words from the same frequency band that fit in the given stem in terms of meaning, the participants who lacked precise knowledge of the word meaning might have had difficulty in choosing the correct answer. However, as the participants went on towards the lower frequency bands, though item difficulty increased, they might have felt more secure to give the correct answer. This behavior may be due to the possibility that the testees might have got used to the format of the test, and more importantly they had to focus solely on one target word form, rather than six, as stated before.

4.4. Correlation of the three test scores with the students' preparatory class exit scores

The fourth search question of the study was as follows:

a. "How well do the VST, VLT and Yes/No Test correlate with one another?"

b. "Which of the VST, VLT and Yes/No Test has the best correlation with the participants' university preparatory class exit scores and yields consistent results?"

The aim of the last research question was first to correlate the students' scores obtained from the three tests' with one another and second to correlate the students' scores obtained from the three tests' with their university preparatory class exit scores. Since the data was not normally distributed and the assumption of normality was not fulfilled, the correlation between the test pairs was conducted with the Spearman's correlation formula and the correlation values were interpreted according to the Figure 14 offered by Mindrila and Balentyne (2013).

Absolute Value of r	Strength of Relationship
r < 0.3	None or very weak
0.3 < r < 0.5	Weak
0.5 < r < 0.7	Moderate
r > 0.7	Strong

Figure 14 Scatterplots and Correlation (Mindrila & Balentyne, 2013)

As seen from Table 14, where the results of the correlation presented, according to Spearman's rho formula, correlation coefficient values between the three test scores were as follows: The Yes/No Test and VLT scores (r=0.59), the Yes/No Test and VST scores (r=0.61) and the VLT and VST scores (r=0.64). These values indicate that there is a positive, linear relationship of moderate strength between each test pair as they are significiant at the 0.01 level (2-tailed). Yet, the overall correlations are not spectacular, according to which the Yes/No Test does not seem to be a very strong predictor of overall VLT or VST performance, and the concurrent validity that has been hoped for between the Yes/No Test and the other two tests is undermined. In other words, the modest correlation between the test pairs suggests that different aspects of vocabulary knowledge are being tested through each test and that the tests are not so strongly connected. This was in contradiction to what was expected as all the three size tests used in this study are claimed in the literature to be measuring the same aspect of word knowledge, namely receptive vocabulary size, and providing the same scores. However, according to the test scores reported before and the correlation values given above, it can be said that though each test measures the form-meaning link, they actually measure different degrees of strength of this aspect.

Table 14

Correlations

			VAR00001	VAR00002	VAR00003	VAR00004
Spearman'srho	VAR00001	CorrelationCoefficient	1,000	,332**	,420**	,314**
		Sig. (2-tailed)	.	,000	,000	,000
		N	372	372	372	372
	VAR00002	CorrelationCoefficient	,332**	1,000	,594**	,615**
		Sig. (2-tailed)	,000	.	,000	,000
		N	372	372	372	372
	VAR00003	CorrelationCoefficient	,420**	,594**	1,000	,644**
		Sig. (2-tailed)	,000	,000	.	,000
		N	372	372	372	372
	VAR00004	CorrelationCoefficient	,314**	,615**	,644**	1,000
		Sig. (2-tailed)	,000	,000	,000	.
		N	372	372	372	372

**. Correlation is significant at the 0.01 level (2-tailed).

Note: V1: Exit Score; V2: the Yes/No Test; V3: the VLT; V4: the VST

Several previous studies showed moderate to strong correlations between overall performance on the Yes/No test and MC tests with independent content (Meara & Buxton, 1987; Shillaw, 1996). The values presented here, though, are lower than the .7 correlation between the Yes/No test results and MC scores reported in Meara and Buxton (1987) and the .84 reported in Anderson and Freebody (1982) for native-speaking children tested on identical items in the two tests. The values are also lower than the .85 correlation for the hits in Mochida and Harrington (2006), where the researchers compared the same item performance employing the Yes/No Test and the VLT, and the correlations of .80 for hits and cfg produced by the minimal instructions as well as the correlations of .73 to .77 for the different formulas produced by the strict instructions between the Yes/No Test and translation performance including the same test items in Eyckmans (2004). In those both studies, test scores proved the Yes/No test to be a robust predictor of VLT performance, which seems to contradict with the results of the present study. The mismatch between this study and the others may be due to a few significant distinctions. For example, whereas Mochida and Harrington and Eyckmans focused on a limited number of students, this study focused on a large number of students (n = 581) from two universities with varying levels of English proficiency. In addition, it was ensured in the present study that all the participants would take totally three size tests, with the same expectations in two different sessions scheduled by the researcher. However, Mochida and Harrington employed the Yes/No test and the VLT, and Eyckmans used the Yes/No test and a translation test. Also, while Mochida and Harrington (2006) focused their analysis on the differences between the frequency level of the word items (2K, 3K, 5K, 10K, or AWL), the analysis of this study focused on the differences between both the frequency levels and the proficiency levels of the participants.

The correlations of the present study are also lower than those found by Stubbe (2012). In that study, 96 real words from the YN forms were tested, and the Pearson correlations of MC scores and YN hits was .79, significant at the .01 level (two-tailed).

On the other hand, the current study presents higher correlations between the Yes/No Test and VLT scores than those found by Cameron (2002). In Cameron’s study, the highest correlation between the Yes/No Test and VLT scores was .45 (Spearman), and none were statistically significant for $p < 0.05$. However, it should not be overlooked that, different from the present study, Cameron’s study did not compare performance on identical items across all frequency levels; it did not include any instruction on re-testing; and it lacked the third measure included here. In this case, since the concurrent validation was based on the correlation between measures with the same content but differing formats, the inferential problems were avoided in the present study. In other words, the opportunity of using test formats with the same content also implied that, in the current study, the correlations should be very high so as to obtain good evidence of concurrent validity of the vocabulary tests as the negative effect of the lack of reliability linked to the inference factor was ruled out. However, the obtained correlations between the vocabulary tests were not very strong, which undermines the expected concurrent validity.

The reason for the lack of the Yes/No Test validity that was established through the low correlations not only between the Yes/No scores and the VST scores but also between the Yes/No scores and the VLT scores might be due to the bias that constituted construct-irrelevant variability. It seemed that several variables that were not related to the construct measured, but instead to the testees' profile, interacted with the lexical knowledge which the test claims to measure. The findings here are thus in line with some earlier work. For instance, Huibregtse et al. (2002) agree that in vocabulary testing a particularly important variable to consider apart from word knowledge is the participants' individual "response style" (Nunnally & Bernstein, 1994) which might have an influence on the participants' responses and affect the scores. Thus, "small differences in response behavior [individual response style] may cause large differences in scores" (Huibregtse et al., 2002, p. 229). It is argued by Eyckmans et al. (2007) as well that there exists a complex interplay among the Yes/No task, learner's profile and particular testing context. Since the factor of self-assessment is part of the Yes/No test format, any individual's self-assessment is not independent of a decision on whether to accept a word as known or not. This decision criterion adopted by the testee might be affected by attitudinal, situational, and motivational factors arising from testing purposes and the test session itself. No matter how rich or partial a student's knowledge of a given word is, s/he is to choose between Yes or No responses, so "[w]hen in doubt, the testee may lean towards either a Yes or a No response simply because of his response style (overestimation, underestimating, refraining from answering) or attitude (analogous to Bourdieu's "economy of practice") (p. 62). In fact, while judging any tested item as known, some learners may be conservative and only check items they are completely sure of, whereas some may feel less rigorous and check items they sense they might know (Read, 2000; Schmitt, 2010), and they, therefore, may show different response styles in the Yes/No Test. In other words, the true underlying knowledge of the two subjects may be the same, but their test scores can be different based on their relative judgement behavior. For example, those learners using an extremely conservative strategy may give a “yes” response to the presented word only when s/he is very confident that it is a known item (Harrington & Carey, 2009). This may result not only in a low FA rate but also in a potential underestimation of the test-taker’s knowledge of real words (Mochida & Harrington, 2006) which might vary noticeably as a function of cultural background and individual differences. Alternatively, when a learner possesses a liberal response strategy, this may result in more “yes” responses to both actual words and pseudowords. Due to the fact that the Yes/No test format seems to elicit a different kind of response behavior in different participants and that self-assessment is unavoidably dependent on one’s personality characteristics, learner attitude is unfortunately hard to control during test administration. It might even depend on linguistic, meta-cognitive or socio-cultural variables and also on the testing context (high- or low-stakes tests). In fact, such factors might have had an influence on the response styles of the participants included in this study and accordingly on their Yes/No Test performance.

Another significant variable, according to Huibregtse et al. (2002), is “guessing”. While choosing from response alternatives, when in doubt, participants try guessing. In general, guesses are not made completely at random. So, the scholars prefer the term “sophisticated guessing”, which means considering the probability of each response alternative, and is used when given a test in which knowledge is not all or nothing but more gradual. That is why it is likely that a good amount of correct guessing could have accounted for the higher scores on the VLT and VST in the present study because at the beginning of these tests, although participants were instructed not to make a guess on the items they do not know the meaning of, they were also instructed to look for the correct alternative when they think they know the word meaning or even part of it. This means, the VST and VLT items might have encouraged the students to apply their partial word knowledge and make informed guesses. However, it is not exactly known how participants react to test instructions and how the guessing factor impacts vocabulary tests. Further study is needed to have a clearer idea about the effects of both factors on the final results of these measures.

When it comes to the correlation value between the VST and VLT, it was not as high as expected, either. It might be because of the fact that the VST and VLT were seemingly similar formats; however, the ways participants were asked to follow while answering the tested items were completely different, which means either test was measuring a different degree of knowledge as supported by the correlation value which was not very strong though the same words were targeted.

In order to answer the second part of the forth research question, it was necessary to correlate the students' scores obtained from the three tests' with their university preparatory class exit scores. As Table 14 shows, the values between the exit scores of the participants and the scores of the each test were as follows: The exit scores and the Yes/No Test scores (r=0.33), the exit scores and the VLT scores (r=0.42) and the exit scores and the VST scores (r=0.31). These values represent a weak yet significant correlation at the 0.01 level (2-tailed).

The value obtained from the correlation of the exit scores and the Yes/No Test scores (r=0.33) of the subjects is lower than the correlations between 0.42 and 0.48, which are significant at p < 0. 00 found by Shillaw (1996). In that study, the researcher correlated the subjects' scores from the Yes/No tests with their total scores on the proficiency test. In another research, it was reported by Meara (1996a) that there is a moderately well correlation between the Yes/No test and other vocabulary tests and tests measuring linguistic skills, especially integrative tests such as the cloze, listening and reading comprehension, where vocabulary knowledge is expected to make an important contribution, though he lacked other measures of vocabulary size to use to cross-validate his checklist tests.

In order to do such cross-validation between different measures, in the current study, the participants' each test score was correlated with their exit score, which must be at least 70 to pass the preparatory class. According to the results, the VLT was the one that had the highest correlation. The reason can be that while in the Yes/No Test and the VST only one target word is tested, in the VLT six words are in the target. In this regard, the VLT seems to be a better tool that can measure an individual's lexical proficiency and also that correlates better with a proficiency exam. Like in a proficiency exam, in the VLT a learner needs to know more words to get a high score.

On the other hand, the reason for the weak correlation between the three vocabulary tests and the exit score in general might be the difference in the content of the vocabulary tests used in this study and the exams the participants take in their institution. In the institutional exams, the participants are not only tested on their recognition based vocabulary knowledge but also on their receptive (reading and listening) and productive (speaking and writing) skills as well as grammar. That is why there appeared a mismatch between the test pairs. Therefore, the weak correlation between the size test scores and the exit scores can actually be attributed to the different constructs underlying each measure. While the size tests measured the recognition knowledge of vocabulary, the university exit score was the total sum of exams measuring different skills and not only the size but also the depth aspect of vocabulary knowledge. Therefore, there is a probability that vocabulary size test scores might not be a reliable indicator of overall language achievement.

Another reason may be the mismatch pertaining to timing of the vocabulary tests and institutional exams the participants took. The three vocabulary tests were given to the students at the beginning of the second term, but the exit scores of the participants represent 50% of their formative in-term exam grades including both terms and 50% of the proficiency exam grades. The high-stakes paper-based proficiency exam takes place at the end of each academic year, which means all the subjects took the proficiency test four months after the administration of the vocabulary tests. In this case, it seems possible that during the four-month period of time, there might have been progress in the participants' vocabulary development as reflected in the weak correlation values between their test scores and exit scores.

Finally, all the findings of the present study mean that as stated by Laufer et al. (2004), though size and strength are related constructs, the participants' performance on the different strength modalities should be reported separately since vocabulary size estimates could be lower or higher depending on the mode the items are measured in. Therefore, for diagnostic purposes, it might be better and more meaningful to separate estimates of size and strength in order to fully understand the overall degree of a learner's knowledge of vocabulary.

This section presented the findings of the study in the light of each research question. It also discussed the relevant findings with reference to the existing evidence offered by studies which have previously been done and also in accordance with any possible factors that might have affected the test scores as well as correlations and led to significant differences. The next chapter will continue with conclusions, recommendations for pedagogical implications, and suggestions for further research.

Chapter V

Conclusion

In this final chapter of the study, in order to outline and emphasize the key findings of the study, some general conclusions are made. After some personal recommendations for pedagogical implications, the researcher concludes the whole thesis with her suggestions for further research.

5.1. Conclusions

In literature, any one of the three tests used in this study is assumed to be a common measure of a learner's receptive vocabulary size, but according to the findings of the current study, these tests do not seem equivalent. The scores obtained from each test are not consistent across proficiency levels, and this means these tests do not measure receptive vocabulary size knowledge in an equal way. The scores of the Y/N Test are lower than those of the VLT and VST at all proficiency levels. That is the Y/N Test is the one which offers conservative estimates and presents low scores. On the other hand, although the VLT and VST scores are close to each other, there is a significant difference between them as the VST presents higher scores and therefore makes higher estimates. In other words, the pattern that is obvious in this study is that the Y/N Test gives the lowest scores, next follows the VLT, and then with the highest scores comes the VST. In this case, it can be said that the Y/N Test format did not actually do what it was claimed to do in literature (measuring the learner's size of the receptive vocabulary knowledge) as effectively as the other two tests.

When the test scores are compared to one another across different frequency levels, it is once again seen that the three vocabulary tests are not consistent. The Y/N Test scores are almost the same as the scores of the VLT solely at the 3K band. At the remaining frequency bands, the participants' Y/N Test scores are significantly low.

According to the scores, there is an inverse relationship pattern between the VLT and VST. At higher frequency levels, the participants' VST scores were lower than their VLT scores. As the participants went over the 3K, 4K, and 5K frequency bands, their scores got higher. The opposite can be said for the VLT. Though participants got higher scores than the VST at higher frequency levels, the scores decreased as participants proceeded towards lower frequency bands. This inverse pattern between the VLT and VST also shows that these two tests are not equivalent measures of vocabulary receptive size.

All these findings mean that the three vocabulary size tests employed in the current study should not be regarded as equivalent measures when they are used in a research study or when they are given to learners as a diagnostic or placement tool to estimate their vocabulary receptive size. Additionally, the differences between these tests should be carefully considered. The reason is that although these three vocabulary tests are assumed to be measuring the size dimension of word knowledge, the test scores in this study clearly show that each measure targets a different type of lexical competente. Therefore, while interpreting the scores obtained from these measures, it does not seem right to say that all of them can offer similar size estimates because of the fact that if a student is not able to recall the meaning of given word form in a Yes/No Test, this does not definitely mean s/he does not know the word as knowledge of form-meaning link is not an all-or-nothing phenomenon. In fact, before a learner reaches the stage of a more advanced component of knowledge like passive recall, s/he may be able to recognize a given word form and even before that, a given word meaning as they are less advanced components of vocabulary knowledge that can be acquired before the recall of meaning aspect. In this case, since knowledge of the form-meaning link depends on what a student is required to do with his/her knowledge of vocabulary, it is not unusual that the tests employed in this study offer different scores and accordingly different knowledge estimates as each of them requires a different degree of knowledge of the same word from the participants.

5.2. Recommendations for Pedagogical Implications

As the results of the current study present, it might be better to use the VLT as a placement and diagnostic tool within an institution. The reason is that the scores obtained from this test correlate better with the students' exit scores.

In any study in which the Yes/No Tests will be employed, the test instructions that will be given to participants should be carefully considered and planned. It should also be kept in mind that it is highly probable that test instructions may have an influence on tested subjects' individual decision making process.

In addition, it should not be forgotten that though there are various tests that can be used in institutions by teachers, they may not be regarded as equally strong measures in terms of their effectiveness in vocabulary testing. In this case, it might be necessary to create awareness in teachers about the availability of such measures and their applicability within the institution. However, theachers should always keep in mind that individiual differences are a big factor in language learning, so it might be even better to give a different test to a different student regarding such differences as they may affect the students' test taking behaviours and may cause the teacher to make wrong assumptions about their language abilities.

Moreover, the teachers, especially the ones in the institution where the researcher has been working, might be encouraged to use these tests in vocabulary assessment. However, they must be aware of the fact that though the tests like the VST and VLT, which have a multiple-choice design, may seem very similar, in fact their underlying constructs are totally different and such differences inarguably affect test scores. Also, regarding the low correlation between the test scores and the exit scores of the participants, it can be offered to teachers to use more contextual measures and instead of just focusing on test scores, they should have a rationale for their test item and design choices.

Furthermore, since in EFL classes vocabulary teaching is usually based on contextualization, in order to help students create better links in their mental lexicons about a new word, teachers should not just focus on the form of the word or its part of speech, but rather they should give information about its meaning, even multiple meanings or other possible word associations, such as collocations. Also, instead of always using the same type of test, they can create tests with different designs to measure the target vocabulary. Having such an approach may help not only students to understand that lexical competence is a multifaceted task but also teachers themselves to realize that strength of knowledge may vary interpersonally; therefore, it might be better to use multiple assessment tools even in the same exam. For example, in order to measure receptive recognition aspect of vocabulary knowledge, instead of giving students a test in which there are just multiple-choice questions, it might be better to give them a test that includes various questions in multiple-choice, matching or fill in the blanks design.

In general, a user of a specific test needs to clearly know what the test is measuring and not measuring. For example, the tests used in this study will probably not function so well when the test user’s aim is measuring vocabulary knowledge needed for productive skills, such as speaking and writing.

Last but not least, researchers and teachers at both Bursa Uludağ University School of Foreign Languages and Bursa Technical University in particular and those who design and administer vocabulary tests in general could benefit from these findings which may give them new insights into the design and implementation of such measures. For instance, as mentioned above, they might consider integrating the three test instruments used in this study to their own testing context as they seem capable of measuring different degrees of vocabulary knowledge strength.

5.3. Suggestions for Further Research

In this study, there were not advanced level subjects. It might be a good idea to do a similar study by including advanced level students in order to see whether differences among test scores would increase or decrease.

In addition, the present study was limited to 1K-5K words, and it did not include the lower word frequency bands, such as 5K-10K. In order to see how scores of the learners will change through these frequency levels, a further study can be carried out.

What is more, so as to obtain better data about the effect of test instructions on participants, studies in which only the Yes/No Tests will be employed and where the specific focus will be on the test instructions are needed. In such studies, it would be easier to check for the potential of the test instructions for influencing the participants' behaviors, such as having a tendency to identify non-words as actual words or being overly conservative in accepting a real word as known.

In order to gather more information about how students' cognitive processes regarding their lexical competence work and how systematic their responses are, more qualitative, in-depth studies which require think-aloud protocols and which focus on participants' rejections are needed in the field.

In this study, one of the research questions was about the relationship between the participants' test scores and their overall language proficiency. It would have been even better to focus on the relationship between the subjects' test scores and their performance regarding the reading comprehension skill as the tests employed here are measures of written receptive vocabulary knowledge, so they could offer higher correlations with a receptive skill like reading, in which vocabulary size is a critical factor.

Also, in order to measure the participants' receptive size knowledge of vocabulary, the participants were provided with direct Turkish equivalents of the target words. Further studies in which subjects are given direct translations of the definitions included in the original tests are needed to see how such a change would affect their test scores.

Lastly, in this study, the participants were not informed about the inclusion of non-words in the Yes/No Test. If they had been given this information, how would the test scores have been affected?

In a nutshell, the primary aim of this study was to cross validate three vocabulary size tests, redesigned in bilingual formats, to see whether they provide equal scores when applied to different level proficiency groups and to show which one can correlate best with an exit score set by the institution. These receptive size tests have different underlying constructs, so it was supported by the findings that they sould not be treated as equal measures of vocabulary knowledge. In addition, if any of them is to be used in a study or institution, it will be better to choose the VLT as it can provide more precise scores and allow for better estimates than the Y/N Test and the VST.

References

Aitchison, J. (2012). *Words in the mind: An introduction to the mental lexicon*. John Wiley & Sons.

Alavi, S. M. (2012). The role of vocabulary size in predicting performance on TOEFL reading item types. *System*, *40*(3), 376-385.

Albrechtsen, D., Haastrup, K., & Henriksen, B. (2008). *Vocabulary and writing in a first and second language: Processes and development*. Basingstoke: Palgrave Macmillan.

Alderson, C. (2001, July). The shape of things to come: Will it be the normal distribution?. In *European language testing in a global context: Proceedings of the ALTE Barcelona conference July* (pp. 1-26).

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment.* London: Continuum.

Alderson, J. C., & Banerjee, J. (2001). Language testing and assessment (Part I). *Language Teaching*, *34*(4), 213-236.

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

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing*, *22*(3), 301-320.

Al-Hazemi, H. (1993). *Low-level EFL vocabulary tests for Arabic speakers.* (Unpublished doctoral dissertation). Available at www.uk.bl.ethos.570245 Dissertations and Theses database.

Alkhofi, A. (2015). *Comparing the receptive vocabulary knowledge of intermediate-level students of different native languages in an intensive English program* (Master's thesis, in the Department of Modern Languages and Literatures in the College of Arts and Humanities at the University of Central Florida. Orlando, Florida).

Alonso, A. C. (2013). Receptive vocabulary size of secondary Spanish EFL learners. *Revista de Lingüística y Lenguas Aplicadas*, *8*(1), 66-75.

Akbarian, I. H. (2010). The relationship between vocabulary size and depth for ESP/EAP learners. *System*, *38*(3), 391-401.

Amirian, S. M. R., Salari, S., Heshmatifar, Z., & Rahimi, J. (2015). A validation study of the newly developed version of vocabulary size test for Persian learners. *International Journal of Education and Research*, *3*(8), 359-380.

Anderson, R. C., & Freebody, P. (1982). Reading comprehension and the assessment and acquisition of word knowledge. *Center for the Study of Reading. Technical Report. No. 249.*

Atkins, A. (2010). Assessing the vocabulary load of text. *Kyoto Sangyo University Journal of Humanities*, *41*(1), 42-51.

Azodi, N., Karimi, F., & Vaezi, R. (2014). Measuring the lexical richness of productive vocabulary in Iranian EFL university students' writing performance. *Theory & Practice in Language Studies*, *4*(9).

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Baddeley, A. D. (1997). *Human memory: Theory and practice*. Psychology Press.

Bardakçi, M. (2016). Breadth and depth of vocabulary knowledge and their effects on L2 vocabulary profiles. *English Language Teaching*, *9*(4), 239-250.

Batista, R. (2014). *A receptive vocabulary knowledge test for French L2 learners with academic reading goals* (Master's thesis, Concordia University. Montreal, Quebec, Canada).

Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M., & Van de Velde, H. (2001). Examining the Yes/No vocabulary test: Some methodological issues in theory and practice. *Language Testing*, *18*(3), 235-274.

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing*, *27*(1), 101-118.

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing*, *16*(2), 131-162.

Bennett, P., & Stoeckel, T. (2013). Developing equivalent forms of a test of general and academic vocabulary. In *JALT 2012 Making a Difference: Conference Proceedings*.

Brown, H. D. (2007). *Teaching by principles: An interactive approach to language pedagogy* (3rd ed.). White Plains, NY: Pearson Education.

Brown, D. (2012). The frequency model of vocabulary learning and Japanese learners. *Vocabulary Learning and Instruction*, *1*(1), 20-28. http://doi.org/10.7820/vli.v01.1.brown

Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language*, *20*(2), 136-163.

Bundgaard-Nielsen, R. L., Best, C. T., & Tyler, M. D. (2011). Vocabulary size is associated with second-language vowel perception performance in adult learners. *Studies in Second Language Acquisition*, *33*(3), 433-461.

Cameron, L. (2002). Measuring vocabulary size in English as an additional language. *Language Teaching Research*, *6*(2), 145-173.

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.

Chen, K. Y. (2011). The impact of EFL students' vocabulary breadth of knowledge on literal reading comprehension. *Asian EFL Journal*, *51*, 30-40.

Chen, T. (2016). *Exploring depth of vocabulary knowledge among CFL learners of higher proficiency levels.* (Unpublished doctoral dissertation). Available at Iowa Research Online: http://ir.uiowa.edu/etd/3057 Dissertations and Theses database.

Choudhury, A. S. (2015). Second/Foreign language lexical competence: Its dimensions and ways of measuring it. *Journal on English Language Teaching*, *5*(3), 34-42.

Chui, S. Y. (2006). A study of the English vocabulary knowledge of university students in Hong Kong. *Asian Journal of English Language Teaching*, *16*, 1-23.

Cobb, T. (1997). Is there any measurable learning from hands-on concordancing?. *System*, *25*(3), 301-315.

Cooper, T. (1997). Assessing vocabulary size: So, what's the problem?, *Language Matters*, 28(1), 96-117. https://doi.org/10.1080/10228199708566122

Corson, D.J. (1995). *Using English words*. Dordrecht: Kluwer Academic Publishers.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213-238.

Coxhead, A., Nation, I. S. P., & Sim, D. (2014). Creating and trialling six forms of the vocabulary size test. *TESOLANZ Journal*, *22*, 13-27.

Coxhead, A., Nation, I. S. P., & Sim, D. (2015). Measuring the vocabulary size of native speakers of English in New Zealand secondary schools. *New Zealand Journal of Educational Studies*, *50*(1), 121-135.

Cronbach, L. J. (1942). An analysis of techniques for diagnostic vocabulary testing. *The Journal of Educational Research*, *36*(3), 206-217.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

Culligan, B. (2015). A comparison of three test formats to assess word difficulty. *Language Testing*, *32*(4), 503-520.

Daller, H., Milton, J., & Treffers-Daller, J. (2007). Editor's introduction. In *Modelling and assessing vocabulary knowledge.* Cambridge: Cambridge University Press.

David, A. (2008). Vocabulary breadth in French L2 learners. *Language Learning Journal*, *36*(2), 167-180.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.

Dewaele, J. M. (2004). Individual differences in the use of colloquial vocabulary: The effects of sociobiographical and psychological factors. *Learning Vocabulary in a Second Language: Selection, Acquisition and Testing*, 127-153.

Dóczi, B., & Kormos, J. (2016). *Longitudinal developments in vocabulary knowledge and lexical organization* (pp. vii+-222). Oxford: Oxford University Press.

Ebel, R. L. & Frisbie, A. D. (1991). *Essentials of educational measuremen*t. 5th Edition. New Delhi: Prentice Hall.

Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, *61*(2), 367-413.

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the vocabulary size test. *Language Testing*, *30*(2), 253-272.

Ellis, N. C. (2002a). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143-188.

Ellis, N. C. (2002b). Reflections on frequency effects in language processing. *Studies in Second Language Acquisition*, *24*(2), 297-339.

Ellis, N., & Beaton, A. (1993). Factors affecting the learning of foreign language vocabulary: Imagery keyword mediators and phonological short-term memory. *The Quarterly Journal of Experimental Psychology*, *46*(3), 533-558. https://doi.org/10.1080/14640749308401062

Elmasry, H. I. (2012). *Depth and breadth of vocabulary knowledge: Assessing their roles in reading comprehension of high-school EFL learners in the UAE* (Doctoral dissertation, The British University in Dubai (BUiD)).

Enayat, M. J., & Amirian, S. M. R. (2016). Vocabulary levels test and word associates test: Can they measure language proficiency?. *International Journal of Assessment and Evaluation in Education,* 6, (32-41). Retrieved from https://www.researchgate.net/publication/309464467

Eyckmans, J. (2000). De Ja/Nee woordenschattoets: Klaar voor gebruik in de klas?. *Toegepaste Taalwetenschap in Artikelen, 64*(1), 117-128. https://doi.org/10.1075/ttwia.64.12

Eyckmans, J. (2004). *Measuring receptive vocabulary size: Reliability and validity of the Yes/No vocabulary test for French-speaking learners of Dutch*. Utrecht: LOT. Retrieved from http://hdl.handle.net/2066/19469

Eyckmans, J., Van de Velde, H., van Hout, R., & Boers, F. (2007). Learners’ response behaviour in Yes/No Vocabulary Tests. In H. Daller, M. Milton, & J. Treffers-Daller

(Eds.), *Modelling and assessing vocabulary knowledge* (pp. 59–76). Cambridge: Cambridge University Press.

Fan, M. (2000). How big is the gap and how to narrow it? An investigation into the active and passive vocabulary knowledge of L2 learners. *RELC Journal*, *31*(2), 105-119.

Fatemipour, H., & Jafari, F. (2015). The effect of dynamic-assessment on the development of passive vocabulary of intermediate EFL learners. *J. Educ. Manage. Stud*, *5*(1), 41-51.

Fulcher, G. (2010). *Practical language testing*. London, UK: Hodder Education.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. New York, NY: Routledge.

Golkar, M., & Yamini, M. (2007). Vocabulary, proficiency and reading comprehension. *The Reading Matrix*, *7*(3).

Greidanus, T., & Nienhuis, L. (2001). Testing the quality of word knowledge in a second language by means of word associations: Types of distractors and types of associations. *The Modern Language Journal*, *85*(4), 567-577.

Greidanus, T., Bogaards, P., van der Linden, E., Nienhuis, L., & de Wolf, T. (2004). The construction and validation of a deep word knowledge test for advanced learners of French. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 191-208). Amsterdam: John Benjamins.

Griffin, G. F. (1992). *Aspects of the psychology of second language vocabulary list learning.* (Doctoral dissertation, University of Warwick). Available at http://go.warwick.ac.uk/wrap/36070 or http://www.uk.bl.ethos.385072 Dissertations and Theses database.

Gyllstad, H. (2004). Testing L2 vocabulary: Current test formats in English as a L2 used at Swedish universities. *The Department of English in Lund: Working Papers in Linguistics*, *4*, 21-40.

Gyllstad, H. (2007). *Testing English collocations: Developing receptive tests for use with advanced Swedish learners*. (Doctoral dissertation, Språk-och litteraturcentrum, Lunds universitet.).

Gyllstad, H. (2009). Designing and evaluating tests of receptive collocation knowledge: COLLEX and COLLMATCH. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 153-170). Basingstoke: Palgrave Macmillan, London.

Gyllstad, H. (2013). Looking at L2 vocabulary knowledge dimensions from an assessment perspective-challenges and potential solutions. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 11-28). Eurosla Monographs Series, 2. EUROSLA.

Gyllstad, H., Vilkaitė, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL-International Journal of Applied Linguistics*, *166*(2), 278-306.

Haastrup, K., & Henriksen, B. (2000). Vocabulary acquisition: Acquiring depth of knowledge through network building. *International Journal of Applied Linguistics, 10*(2), 221-240.

Harrington, M., & Carey, M. (2009). The on-line Yes/No test as a placement tool. *System*, *37*(4), 614-626.

Harsch, C., & Hartig, J. (2015). Comparing C-tests and Yes/No vocabulary size tests as predictors of receptive language skills. *Language Testing*, 1-21.

Hasan, K., & Shabdin, A. A. (2016). Conceptualization of depth of vocabulary knowledge with academic reading Comprehension. *PASAA: Journal of Language Teaching and Learning in Thailand*, *51*, 235-268.

Hashimoto, B. J. (2016). *Rethinking vocabulary size tests: Frequency versus item difficulty.* (Master's thesis, Department of Linguistics and English Language Brigham Young University). Available at https://scholarsarchive.byu.edu/etd/5958 Dissertations and Theses database.

Hatami, S., & Tavakoli, M. (2012). The role of depth versus breadth of vocabulary knowledge in success and ease in L2 lexical inferencing. *TESL Canada Journal*, 1-1.

Hayashi, Y., & Murphy, V. (2011). An investigation of morphological awareness in Japanese learners of English. *Language Learning Journal*, *39*(1), 105-120. https://doi.org/10.1080/09571731003663614

Hazenberg, S., & Hulstun, J. H. (1996). Defining a minimal receptive second-language vocabulary for non-native university students: An empirical investigation. *Applied Linguistics*, *17*(2), 145-163.

Heilman, M., & Eskenazi, M. (2008, June). Self-assessment in vocabulary tutoring. In *International Conference on Intelligent Tutoring Systems* (pp. 656-658). Springer, Berlin, Heidelberg.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, *21*(2), 303-317.

Henriksen, B. (2006). Exploring the quality of lexical knowledge in the language learner's L1 and L2. *AFinLAn vuosikirja*.

Henriksen, B. (2008). Declarative lexical knowledge. In Albrechtsen, D., Haastrup, K., and Henriksen, B., *Vocabulary and writing in a first and second language*. Basingstoke: Palgrave Macmillan.

Henriksen, B. (2013). Research on L2 learners' collocational competence and development-a progress report. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary* *acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 29–56). Eurosla Monographs Series, 2. EUROSLA.

Hirsh, D. (2015). Researching vocabulary. In B. Paltridge & A. Phakiti (Eds.), *Research methods in applied linguistics: A practical resource* (pp. 369-385). Bloomsbury Publishing.

Hirsh, D., & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure?. *Reading in a Foreign Language*, *8*, 689-689.

Horst, M., & Collins, L. (2006). From faible to strong: How does their vocabulary grow?. *Canadian Modern Language Review*, *63*(1), 83-106.

Hsueh-chao, M. H., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), 403-430. Retrieved from http://nflrc.hawaii.edu/rfl/PastIssues/rfl131hsuehchao.pdf

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). Available at https://www.proquest.com Dissertations and Theses database.

Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes-no vocabulary test: Correction for guessing and response style. *Language Testing*, *19*(3), 227-245.

Jarema, G., & Libben, G. (2007). Introduction: Matters of definition and core perspectives. In G. Jarema and G. Libben (Eds.), *The mental lexicon: Core perspectives* (pp. 1–5). Oxford: Elsevier Press.

Jordan, E. (2013). A true/false translation test for English vocabulary size assessment in a Japanese context. *Polyglossia*, *25*, 27-45.

Karami, H. (2012). The development and validation of a bilingual version of the Vocabulary Size Test. *RELC Journal*, *43*(1), 53-67.

Kempe, V., & MacWhinney, B. (1996). The crosslinguistic assessment of foreign language vocabulary learning. *Applied Psycholinguistics*, *17*(2), 149-183.

Kremmel, B., & Schmitt, N. (2018). Vocabulary levels test. *The TESOL Encyclopedia of English Language Teaching*, 1-7.

Koda, K. (1996). L2 word recognition research: A critical review. *The Modern Language Journal*, *80*(4), 450-460.

Koya, T. (2005). *The acquisition of basic collocations by Japanese learners of English*. (Unpublished doctoral dissertation). Available at http://dspace.wul.waseda.ac.jp/dspace/bitstream/2065/5285/3/Honbun-4160.pdf Dissertations and Theses database.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. Arnaud & H. Benoit (Eds.), *Vocabulary and applied linguistics* (pp. 126–32). London: Palgrave Macmillan.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different?. *Applied Linguistics*, *19*(2), 255-271.

Laufer, B., & Aviad–Levitzky, T. A. M. I. (2017). What type of vocabulary knowledge predicts reading comprehension: Word meaning recall or word meaning recognition?. *The Modern Language Journal*, *101*(4), 729-741.

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/10.1191/0265532204lt277oa

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399-436. https://doi.org/10.1111/j.0023- 8333.2004.00260.x

Laufer, B., & McLean, S. (2016). Loanwords and vocabulary size test scores: A case of different estimates for different L1 learners. *Language Assessment Quarterly*, *13*(3), 202-217. https://doi.org/10.1080/15434303.2016.1210611

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 33-51. https://doi.org/10.1191/026553299672614616

Laufer, B.,& Paribakht, T. S. (1998). The relationship between passive and active vocabularies: Effects of languagelearning context. *Language Learning*, *48*(3), 365-391.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*(1), 15-30.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners’ English. *Language Learning*, *61*(2), 647-672.

Lee, S. (2012). Receptive vocabulary size for reading English and American novels. *Multimedia-Assisted Language Learning*, *15*(4), 103-125.

Lee, S. H., & Muncie, J. (2006). From receptive to productive: Improving ESL learners' use of vocabulary in a postreading composition task. *Tesol Quarterly*, *40*(2), 295-320.

Lemmouh, Z. (2010). *The relationship among vocabulary knowledge, academic achievement and the lexical richness in writing in Swedish university students of English* (Doctoral dissertation, Department of English, Stockholm University).

Levitzky-Aviad, T., & Laufer, B. (2013). Lexical properties in the writing of foreign language learners over eight years of study: Single words and collocations. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 127-148). Eurosla Monographs Series, 2. EUROSLA.

Li, M., & Kirby, J. R. (2014). The effects of vocabulary breadth and depth on English reading. *Applied Linguistics*, *36*(5), 611-634.

Li, L., & MacGregor, L. J. (2010). Investigating the receptive vocabulary size of university-level Chinese learners of English: How suitable is the vocabulary levels test?. *Language and Education*, *24*(3), 239-249.

Lin, L. H., & Morrison, B. (2010). The impact of the medium of instruction in Hong Kong secondary schools on tertiary students' vocabulary. *Journal of English for Academic Purposes*, *9*(4), 255-266.

Ling, G. (2015). Predictability of vocabulary size on learners' EFL proficiency: Taking VST, CET4 and CET6 as instruments. *Studies in Literature and Language*, *10*(3), 18.

Liu, J. (2016). Role of vocabulary levels test (VLT) in Chinese undergraduate students' comprehension of textbooks. *Journal of Language Teaching and Research*, *7*(2), 364-369.

Makarchuk, D. (2013). University freshmen's EFL receptive and productive recall vocabulary knowledge and use. *English Teaching*, *68*(4), 217-239.

Marzban, A., & Hadipour, R. (2012). Depth versus breadth of vocabulary knowledge: Assessing their roles in Iranian intermediate EFL students' lexical inferencing success through reading. *Procedia-Social and Behavioral Sciences*, *46*, 5296-5300.

Maskor, Z. M., & Baharudin, H. (2016). Receptive vocabulary knowledge or productive vocabulary knowledge in writing skill, which one important?. *International Journal of Academic Research in Business and Social Sciences*, *6*(11), 2222-6990.

McLean, S., & Kramer, B. (2015). The creation of a new vocabulary levels test. *Shiken*, *19*(2), 1-11.

McLean, S., & Kramer, B. (2016). The development of a Japanese bilingual version of the new vocabulary levels test. *VERB* 5(1), 2-5).

McLean, S., Kramer, B., & Beglar, D. (2015a). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*, *19*(6), 741-760. http://doi.org/10.1177/1362168814567889

McLean, S., Kramer, B., & Stewart, J. (2015b). An empirical examination of the effect of guessing on vocabulary size test scores. *Vocabulary Learning and Instruction*, *4*(1), 26-35. https://doi.org/10.7820/vli.v04.1.mclean.et.al

Meara, P. (1990). A note on passive vocabulary. *Second Language Research*, *6*(2), 150-154. Retrieved from http://www.jstor.org/stable/43104407

Meara, P. (1992). EVST (Eurocentre) Eng. [online]. Available at http://www.lextutor.ca/tests/

Meara, P. (1993). The bilingual lexicon and the teaching of vocabulary. In R. Schreuder & B. Weltens (Eds.), *The bilingual lexicon* (pp. 279–297). Amsterdam/Philadelphia: John Benjamins.

Meara, P. (1994). The complexities of simple vocabulary tests. Retrieved from www.lognostics.co.uk/vlibrary/index.htm

Meara, P. (1996a). The dimensions of lexical competence. Retrieved from www.lognostics.co.uk/vlibrary/index.htm

Meara, P. (1996b). The vocabulary knowledge framework. *Vocabulary Acquisition Research Group Virtual Library*. Retrieved from www.lognostics.co.uk/vlibrary/index.htm

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.

Meara, P. M. (2005a). X_Lex: the Swansea vocabulary levels test. v2.05. Swansea: Lognostics.

Meara, P. M. (2005b). Lex vocabulary tests v2.0. Swansea: University of Wales, Centre for Applied Language Studies.

Meara, P. (2010). EFL vocabulary tests. Swansea: _lognostics second edition 2010.

Retrieved from http://www.lognostics.co.uk/vlibrary/meara1992z.pdf

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*(2), 142-154.

Meara, P., & Fitzpatrick, T. (2000). Lex30: An improved method of assessing productive vocabulary in an L2. *System*, *28*(1), 19-30.

Meara, P., & Jones, G. (1988). Vocabulary size as a placement indicator. Retrieved from www.lognostics.co.uk/vlibrary/index.htm

Meara, P., Lightbown, P. M., & Halter, R. H. (1994). The effect of cognates on the applicability of YES/NO vocabulary tests. *Canadian Modern Language Review*, *50*(2), 296-311.

Meara, P. M., & Milton, J. L. (2005). X_Lex: the Swansea vocabulary levels test. v2. 05. *Swansea: Lognostics*. Retrieved from www.lognostics.co.uk/vlibrary/index.htm

Meara, P. M., & I. Miralpeix. (2006). Y_Lex: The Swansea advanced vocabulary levels test. v2.05. *Swansea: Lognostics*. Retrieved from www.lognostics.co.uk/vlibrary/index.htm

Meara , P. M., & Wolter , B . (2004). V_Links: Beyond vocabulary depth . *Angles on the English- Speaking World, 4,* 85-96. Retrieved from www.lognostics.co.uk/vlibrary/index.htm

Melka, F. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, and pedagogy* (pp. 84–102). Cambridge: Cambridge University Press.

Melka Teichroew, F.J. (1982). Receptive versus productive vocabulary: A survey. *Interlanguage Studies Bulletin*, 5-33.

Milton, J. (2007). Lexical profiles, learning styles and the construct validity of lexical size tests. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledg*e (pp. 47–58). Cambridge: CUP.

Milton, J. (2009). *Measuring second language vocabulary acquisition* (Vol. 45). Bristol, UK: Multilingual Matters.

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, 211-232.

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 57-78). Eurosla Monographs Series, 2. EUROSLA.

Milton, J., & Alexiou, T. (2009). Vocabulary size and the Common European Framework of Reference for languages. In B. Richards, H.M. Daller, D. D. Malvern, P. Meara, J. Milton, & J. Treffers-Daller (Eds.), *Vocabulary studies in first and second language acquisition: The interface between theory and application* (pp. 194-211). Basingstoke, UK: Palgrave Macmillan.

Milton, J., & Fitzpatrick, T. (Eds.). (2014). *Dimensions of vocabulary knowledge.* Basingstoke, UK: Palgrave Macmillan.

Milton, J., & N. Hopkins. (2005). *Aural Lex.* Swansea, UK: Swansea University.

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in English as a foreign language. In R. Chac´on-Beltr´an, C. Abello-Contesse, & M. Torreblanca-L´opez (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 83–98). Bristol, UK: Multilingual Matters.

Mindrila, D., & Balentyne, P. (2013). *Scatterplots and correlations.* Retrieved from https://www.westga.edu/academics/research/vrc/assets/docs/scatterplots_and_correlation_notes.pdf

Mizumoto, A. (2012). Exploring the effects of self-efficacy on vocabulary learning strategies. *SiSAL Journal, 3*(4), 423–437. Retrieved from http://sisaljournal.org/archives/dec12/mizumoto/

Mochida, K., & Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. *Language Testing*, *23*(1), 73-98.

Mochizuki, M. (2012). Four empirical vocabulary test studies in the three dimensional framework. *Vocabulary Learning and Instruction*, *1*(1), 44-52.

Mondria, J. A., & Wiersma, B. (2004). Receptive, productive, and receptive+ productive L2 vocabulary learning: What difference does it make? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing 15*(1), (pp. 79-100). Amsterdam: John Benjamins.

Mutlu, G., & Kaşlıoğlu, Ö. (2016). Vocabulary size and collocational knowledge of Turkish EFL learners. *Journal of Theory and Practice in Education*, 12(6), 1231-1252

Nadarajan, S. (2008). Assessing in-depth vocabulary ability of adult ESL learners. *The International Journal of Language Society and Culture*, *26*, 93-106.

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), 387-401.

Nation, I. S. P. (1983) Testing and teaching vocabulary. Guidelines 5(1), 12-25. Retrieved from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/1983-Testing-and-teaching.pdf

Nation, I. S. P. 1990. *Teaching and learning vocabulary*. Boston, MA: Heinle & Heinle.

Nation, I. S. P. (2000). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

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

Nation, I. (2006). How large a vocabulary is needed for reading and listening?. *Canadian Modern Language Review*, *63*(1), 59-82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P. (2008). *Teaching vocabulary: Strategies and techniques.* Boston: Heinle.

Nation, I. S. P. (2012a, August). *Measuring vocabulary size in an uncommonly taught language.* In International Conference on Language Proficiency Testing in the Less Commonly Taught Languages (pp. 17-18).

Nation, I. S. P. (2012b, October 23). The vocabulary size test. Retrieved from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Vocabulary- Size-Test-information-and-specifications.pdf

Nation, I. S. P., & Beglar, D. (2007a). A vocabulary size test. *The Language Teacher 31 (7)*, 9-13. Retrieved from http://jalt-publications.org/tlt/issues/2007-07_31.7

Nation, I. S. P., & Beglar, D. (2007b). Vocabulary size test. University of Wellington. http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx

Nation, P., & Chung, T. (2009). Teaching and testing vocabulary. In M.H. Long & C.J. Doughty (Eds.), *The handbook of language teaching* (pp. 543-559). Malden: Wiley-Blackwell.

Nation, P., & Coxhead, A. (2014). Vocabulary size research at Victoria University of Wellington, New Zealand. *Language Teaching*, *47*(3), 398-403.

Nation, P., & Gu, P.Y. (2007). *Focus on vocabulary*. Sydney, Australia: National Centre for English Language Teaching and Research, Macquarie University.

Nation, P., & Meara, P. (2010). Vocabulary. In N. Schmitt (Ed.), *An introduction to applied linguistics (2nd Edition)* (pp. 34-52). Hodder Education, UK: London.

Nemati, A. (2010). Active and passive vocabulary knowledge: The effect of years of instruction. *The Asian EFL Journal Quarterly*, *12*(1), 30-46.

Nergis, A. (2013). Exploring the factors that affect reading comprehension of EAP learners. *Journal of English for Academic Purposes*, *12*(1), 1-9.

Nizonkiza, D., & Van den Berg, K. (2014). The dimensional approach to vocabulary testing: What can we learn from past and present practices?. *Stellenbosch Papers in Linguistics*, *43*, 45-61.

Nguyen, L.T.C., & Nation, I.S.P. (2011). A bilingual size test of English for Vietnamese learners. *RELC Journal*, *42*(1), 86–99. https://doi.org/10.1177/0033688210390264

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory. 3rd edition.* New York: McGraw-Hill.

Ozturk, M. (2003, September). *Lexical competence in the Common European Framework of Reference for languages.* In I. International Symposium on the Common European Framework and Foreign Language Education in Turkey, Bursa.

Ozturk, M. (2007). Multiple-choice test items of foreign language vocabulary. *Uludağ Üniversitesi Eğitim Fakültesi Dergisi*, *20*(2), 399-426.

Ozturk, M. (2013). L2 Vocabulary growth in intensive language study: The effects of proficiency level and word frequency. Retrieved from http://researchgate.net

Ozturk, M. (2015). Vocabulary growth of the advanced EFL learner. *The Language Learning Journal*, *43*(1), 94-109. doi: 10.1080/09571736.2012.708053. http://dx.doi.org/10.1080/09571736.2012.708053

Ozturk, M. (2016). Second language vocabulary growth at advanced level. *The Language Learning Journal*, *44*(1), 6-16. doi: 10.1080/09571736.2012.708054. http://dx.doi.org/10.1080/09571736.2012.708054

Palmer, H.E. (1917) *The scientific study and teaching of languages.* London: Harrap.

Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring Yes–No vocabulary tests: Reaction time vs. nonword approaches. *Language Testing*, *29*(4), 489-509.

Pignot-Shahov, V. (2012). Measuring L2 receptive and productive vocabulary knowledge. *Language Studies Working Papers*, *4*(1), 37-45.

Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review*, *56*(2), 282-308.

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

Qian, D. D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing*, *21*(1), 28-52.

Read, J. (1988). Measuring the vocabulary knowledge of second langauge learners. *RELC Journal*, *19*(2), 12-25.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*(3), 355-371.

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

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language*: *Selection, acquisition, and testing* (pp. 209–227). Amsterdam: John Benjamins.

Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies*, *7*(2), 105-126.

Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing*, *18*(1), 1-32.

Read, J., & Nation, P. (1986). *Some issues in the testing of vocabulary knowledge*. Paper presented at a Language Testing Symposium in Honor of John B. Carroll and Robert Lado (Quiryat Anavim, Israel, May 11-13, 1986).

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

Richards, B., Malvern, D., & Graham, S. (2008). Word frequency and trends in the development of French vocabulary in lower-intermediate students during Year 12 in English schools. *Language Learning Journal*, *36*(2), 199-213.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing*, *15*(1), 1-20.

Ruegg, R. (2007). The English vocabulary level of Japanese junior high school students. In K. Bradford Watts, T. Muller, & M. Swanson (Eds.), *JALT2007 Conference Proceedings, 103-109. Tokyo: JALT*. Retrieved from http://jalt-publications.org/archive/proceedings/2007/E013.pdf

Sakai, R. (2009). *Receptive and productive vocabulary size of Japanese undergraduate students: A correlation with reading and writing proficiency.* (Master's thesis, Teaching English as a Foreign Language. Bangkok: Graduate School, Srinakharinwirot University).

Saville-Troike, M. (1984). What really matters in second language learning for academic achievement?. *TESOL Quarterly*, *18*(2), 199-219.

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

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

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329-363.

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

Schmitt, N. (2014). Conceptual review article size and depth of vocabulary knowledge: what the research shows. *Language Learning, 64*(4), 913-951.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(1), 26-43.

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition*, *19*(1), 17-36.

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*(4), 484-503.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language Testing*, *18*(1), 55-88. https://doi.org/10.1177/026553220101800103

Shah, S. K., Gill, A. A., Mahmood, R., & Bilal, M. (2013). Lexical richness, a reliable measure of intermediate L2 learners' current status of acquisition of English language. *Journal of Education and Practice*, *4*(6), 42-47.

Shen, Z. (2008). The roles of depth and breadth of vocabulary knowledge in EFL reading performance. *Asian Social Science*, *4*(12), 135-137.

Shillaw, J. (1996). The application of Rasch modeling to yes/no vocabulary tests. *Vocabulary Acquisition Research Group*. Retrieved from www.lognostics.co.uk/vlibrary/index.htm

Shillaw, J. (1999). *The application of the Rasch model to Yes/No vocabularly tests*. (Unpublished doctoral dissertation). Available at http://www.uk.bl.ethos.594177 Dissertations and Theses database.

Shin, D., Chon, Y. V., & Kim, H. (2011). Receptive and productive vocabulary sizes of high school learners: What next for the basic word list. *English Teaching*, *66*(3), 123-148.

Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing*, *24*(1), 99-128.

Sims, V. M. (1929). The reliability and validity of four types of vocabulary tests. *The Journal of Educational Research*, *20*(2), 91-96.

Sonbul, S., & Schmitt, N. (2009). Direct teaching of vocabulary after reading: Is it worth the effort?. *ELT Journal*, *64*(3), 253-260.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*(2), 139-152.

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, *31*(4), 577-607.

Stewart, J. (2012). A multiple-choice test of active vocabulary knowledge. *Vocabulary Learning and Instruction*, *1*(1), 53-59. http://dx.doi.org/10.7820/vli.v01.1.stewart

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST?. *Language Assessment Quarterly*, *11*(3), 271-282.

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.

Stoeckel, T., & Bennett, P. (2015). A test of the new General Service List. *Vocabulary Learning and Instruction*, *4*(1), 1-8.

Stubbe, R. (2012). Do pseudoword false alarm rates and overestimation rates in Yes/No vocabulary tests change with Japanese university students' English ability levels?. *Language Testing*, *29*(4), 471-488.

Stubbe, R. (2014). Do Japanese students overestimate or underestimate their knowledge of English loanwords more than non-loanwords on yes-no vocabulary tests. *Vocabulary Learning and Instruction*, *3*(1), 29-43. Advance online publication. https://doi.org/10.7820/vli.v03.1.stubbe

Stubbe, R., & Stewart, J. (2012). Optimizing scoring formulas for yes/no vocabulary tests with linear models. *Shiken Research Bulletin*, *16*(2), 2-7. Retrieved from http://teval.jalt.org/node/12

Stubbe, R., Stewart, J., & Pritchard, T. (2010). Examining the effects of pseudowords in yes/no vocabulary tests for low level learners. *Kyushu Sangyo University Language Education and Research Center Journal*, *5*, 5-23.

Stubbe, R., O'Sullivan, C., Boston, J., Porter, M., Grumbine, R., & Latz, D. (2011). Who check more pseudowords, low-level or high-level students? In A. Stewart (Ed.), *JALT2010 Conference Proceedings*. Tokyo: JALT. Retrieved from http://jalt2010proc-76.pdf

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 471–83). Mahwah, NJ: Lawrence Erlbaum.

Takala, S. (1985). Estimating students' vocabulary sizes in foreign language teaching. *AFinLAn Vuosikirja*, 157-165.

Tanaka, S. (2012). New directions in L2 lexical development. *Vocabulary Learning and Instruction*, *1*(1), 1-9. http://dx.doi.org/10.7820/vli.v01.1.tanaka

Tschirner, E. (2004). Breadth of vocabulary and advanced English study: An empirical investigation. *Electronic Journal of Foreign Language Teaching*, *1*(1), 27-39.

Tseng, W. T., & Schmitt, N. (2008). Toward a model of motivated vocabulary learning: A structural equation modeling approach. *Language Learning*, *58*(2), 357-400.

Uden, J., Schmitt, D., & Schmitt, N. (2014). Jumping from the highest graded readers to ungraded novels: Four case studies. *Reading in a Foreign Language*, *26*(1), 1-28.

Uzun, L., & Salihoglu, U. (2009). English-Turkish cognates and false cognates: Compiling a corpus and testing how they are translated by computer programs. *Poznań Studies in Contemporary Linguistics*, *45*(4), 569-593.

van Hout, R., & Vermeer, A. (2007). Comparing measures of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 93–115). Cambridge: Cambridge University Press.

Varnaseri, M., & Farvardin, M. T. (2016). The relationship between depth and breadth of vocabulary knowledge and writing performance of Iranian MA students of TEFL. *Modern Journal of Language Teaching Methods*, *6*(2), 544.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics*, *22*(2), 217-234.

Verspoor, M., & Cremer, M. (2008). Research on foreign-language teaching and learning in the Netherlands (2002–2006). *Language Teaching*, *41*(2), 183-211.

Wang, Y., & Du, W. (2014). Study on the validity of bilingual Mandarin version of vocabulary size test. *International Journal of English Linguistics*, *4*(6), 113.

Waring, R. (1997). A comparison of the receptive and productive vocabulary sizes of some second language learners. *Immaculata,* (1), 53-68.

Waring, R. (1998). Receptive and productive foreign language vocabulary size II. Unpublished manuscript. Available at http://www.l.harenet.ne.jp/-war-ing/vocabindex.html

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader. *Reading in a Foreign Language*, *15*(2), 130–163.

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, *30*(1), 79-95.

Webb, S. (2009). The effects of receptive and productive learning of word pairs on vocabulary knowledge. *RELC Journal*, *40*(3), 360-376.

Webb, S. A., & Sasao, Y. (2013). New directions in vocabulary testing. *RELC Journal*, *44*(3), 263-277.

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review*, *53*(1), 13-40.

Wolter, B. (2001). Comparing the L1 and L2 mental lexicon: A depth of individual word knowledge model. *Studies in Second Language Acquisition*, *23*(1), 41-69.

Xing, P., & Fulcher, G. (2007). Reliability assessment for two versions of vocabulary levels tests. *System*, *35*(2), 182-191.

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication*, *3*(2), 215-229.

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.

Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition*, *27*(4), 567-595.

Zhang, X. (2013). The I don't know option in the vocabulary size test. *TESOL Quarterly*, *47*(4), 790-811.

Zhang, X., Pan, Y., & Xu, X. (2014). The correlation study on the relationship between the depth of vocabulary knowledge and comprehensive ability. *Studies in Literature and Language*, *9*(1), 94.

Zhao, P., & Ji, X. (2018). Validation of the Mandarin version of the vocabulary size test. *RELC Journal*, *49*(3), 308-321.

Zhong, H. (2011). *Learning a word: From receptive to productive vocabulary use.* In the *Asian Conference on Language Learning: Official Conference Proceedings* (pp. 116-126).

Zhong, H. F. (2018). The relationship between receptive and productive vocabulary knowledge: a perspective from vocabulary use in sentence writing. *The Language Learning Journal*, *46*(4), 357-370. https://doi.org/10.1080/09571736.2015.1127403

Zhong, H., & Hirsh, D., (2009). Vocabulary growth in an English as a foreign language context. *University of Sydney Papers in TESOL*, *4*(4), 85-113.

Zhou, S. (2010). Comparing receptive and productive academic vocabulary knowledge of Chinese EFL learners. *Asian Social Science*, *6*(10), 14.

Zimmerman, K. J. (2004). *The role of vocabulary size in assessing second language proficiency* (Master's thesis, Department of Linguistics Brigham Young University). Available at http://scholarsarchive.byu.edu/etd/578

Zimmerman, J., Broder, P. K., Shaughnessy, J. J., & Underwood, B. J. (1977). A recognition test of vocabulary using signal-detection measures, and some correlates of word and nonword recognition. *Intelligence*, *1*(1), 5-31.

List of Appendices

Appendix 1: Receptive Size Tests

Size Test

07.01.2016

Ad-Soyad: ________________

Öğrenci No: _______________

Sınıf: ____________________

İmza: ____________________

Bu testlerin sonuçlarından elde edilecek veriler, bir yüksek lisans tez çalışmasına kaynak oluşturacaktır.

Bu bir **İngilizce sözcük bilgisi** testidir.

Lütfen **koyu** olarak yazılmış ve **örnek cümlede** kullanılmış **İngilizce hedef kelimenin Türkçe karşılığını** veren seçeneği (a, b, c veya d) işaretleyiniz.

Örnek Soru

game: I like this **game**.

a. yiyecek

b. hikaye

c. oyun Doğru cevap **c** seçeneğidir.

d. insan

Eğer kelimenin anlamını bilmiyorsanız, lütfen o soruyu boş bırakınız ve diğer soruya geçiniz.

Eğer, sözcüğün anlamı hakkında herhangi bir tahmininiz var ise, soruya yanıt vermeye çalışınız.

Katılımınız için çok teşekkür ederim.

Teste geçebilirsiniz. BAŞARILAR…

Okt. Sezen AKSU BALAMUR

1. pressure: They used too much <pressure>.
a. pastırma
b. tereyağı
c. basınç
d. nakit para

2. eclipse: There was an <eclipse>.
a. yardımcı
b. karınca
c. koridor
d. güneş tutulması

3. joke: We did not understand his <joke>.
a. espri
b. hikaye
c. konuşma
d. fikir

4. crook: They were <crooks>.
a. hizmetçi
b. sahtekar
c. şövalye
d. göçmen

5. flower: He gave me a <flower>.
a. elbise
b. çiçek
c. saat
d. anahtar

6. flock: Look at this <flock>.
a. sürü
b. mezarlık
c. sera
d. yara

7. see: They <saw> it.
a. sıkıca kapatmak
b. beklemek
c. görmek
d. çalıştırmak

8. dig: Our dog often <digs>.
a. oyun oynamak
b. yeri kazmak
c. uykusu gelmek
d. suya girmek

9. resist: People naturally <resist> change.
a. hoş karşılamak
b. talep etmek
c. yaratmak
d. direnmek

10. weave: She <weaves> these things.
a. sınıflandırmak
b. dokumak
c. etiketlemek
d. ithal etmek

11. poor: We are <poor>.
a. fakir
b. mutlu
c. ilgili
d. güçlü

12. latter: The man dies in the <latter> part of the story.
a. başlangıç
b. son
c. en önemli
d. ikinci

13. patience: He has no <patience>.
a. merhamet
b. ateş
c. sabır
d. önyargı

14. reward: He likes <rewards>.
a. deney
b. roman
c. ödül
d. anket

15. strap: She changed the <strap>.
a. kayış
b. yastık
c. kilim
d. kaşık

16. grow: Children <grow> fast.
a. resim çizmek
b. konuşmak
c. koşmak
d. büyümek

17. rove: They are <roving>.
a. münakaşa etmek
b. dolaşmak
c. mırıldanmak
d. gevezelik etmek

18. lonesome: He felt <lonesome>.
a. yalnız
b. gergin
c. halsiz
d. umutsuz

19. crab: Do you like <crabs>?
a. kahvaltılık gevrek
b. sirke
c. kiraz
d. yengeç

20. ambition: He has no <ambition>.
a. uzmanlık
b. hırs
c. boş vakit
d. şöhret

21. slaughter: We read about the <slaughter> in the paper.
a. kuraklık
b. kuşatma
c. katliam
d. deprem

22. deficit: The company had a large <deficit>.
a. hesap açığı
b. yatırım
c. indirim
d. etki

23. basis: This was used as the < basis>.
a. çöp kutusu
b. temel
c. değirmen
d. askı

24. time: They have a lot of <time>.
a. para
b. yiyecek
c. zaman
d. tecrübe

25. lend: She often <lends> her books.
a. ödünç vermek
b. resimlerle süslemek
c. düzenlemek
d. tavsiye etmek

26. envy: They <envy> each other.
a. sarılmak
b. taklit etmek
c. küçük düşürmek
d. imrenmek

27. seal: They <sealed> it.
a. onarmak
b. mühürlemek
c. geliştirmek
d. planlamak

28. allege: He <alleged> it.
a. ele geçirmek
b. icat etmek
c. iddia etmek
d. değerlendirmek

29. delicious: The meal was really <delicious>.
a. lezzetli
b. pahalı
c. berbat
d. özel

30. fragile: It is very <fragile>.
a. bol
b. tuhaf
c. kıymetli
d. kırılabilir

31. result: They were waiting for the <results>.
a. ürün
b. yorum
c. müşteri
d. sonuç

32. speech: We enjoyed the <speech>.
a. konuşma
b. yarışma
c. mobilya
d. yolculuk

33. remedy: We found a good <remedy>.
a. kaynak
b. çare
c. aday
d. tamirci

34. refuse: She <refused> the invitation.
a. minnetkar olmak
b. görmezden gelmek
c. reddetmek
d. almak

35. drive: He <drives> fast.
a. yüzmek
b. öğrenmek
c. konuşmak
d. araba kullanmak

36. versatile: He is a <versatile> person.
a. sevilen
b. düzenli
c. çok yönlü
d. inatçı

37. compound: They made a new <compound>.
a. rica
b. bileşik
c. tahmin
d. cihaz

38. plant: There are many different <plants> here.
a. bitki
b. ev
c. insan
d. hayvan

39. peasant: He did a lot for the <peasants>.
a. köylü
b. mülteci
c. ergen
d. muhafazakar

40. knee: Take care of your <knee>.
a. bisiklet
b. diz
c. evcil hayvan
d. oyuncak

41. solution: There is no <solution>.
a. zafer
b. tutku
c. ekonomik durgunluk
d. çözüm

42. marble: It was made of <marble>.
a. kil
b. un
c. mermer
d. çadır bezi

43. applaud: All the children <applauded>.
a. alkışlamak
b. göz kırpmak
c. kızarmak
d. kıkır kıkır gülmek

44. warn: They <warned> him.
a. tanımak
b. etkilemek
c. uyarmak
d. seçmek

45. pave: They <paved> it.
a. çözmek
b. tercüme etmek
c. güncelleştirmek
d. taş döşemek

46. diminish: It has <diminished>.
a. yola çıkmak
b. azalmak
c. kararmak
d. bozulmak

47. competent: She was <competent>.
a. sadık
b. istekli
c. yetkin
d. savunmasız

48. dubious: They were <dubious>.
a. kuşkulu
b. güzel kokulu
c. nesli tükenmiş
d. önemsiz

49. round: It is <round>.
a. kocaman
b. kısa
c. kalın
d. yuvarlak

50. awe: They felt <awe>.
a. minnetkarlık
b. üzüntü
c. hayranlık
d. dehşet

51. behavior: Look at her <behavior>.
a. seyirci
b. davranış
c. rakip
d. bütçe

52. danger: The children are still in <danger>.
a. hayret
b. tehlike
c. korku
d. ağrı

53. maintain: Can they <maintain> it?
a. sürdürmek
b. teslim etmek
c. tarif etmek
d. ispat etmek

54. weep: He <wept>.
a. yok olmak
b. göz kırpmak
c. ağlamak
d. övünmek

55. threshold: They raised the <threshold>.
a. tabut
b. merdiven
c. eşik
d. bilet ücreti

56. drawer: The <drawer> was empty.
a. çekmece
b. otoyol şeridi
c. buzdolabı
d. akü

57. fracture: They found a <fracture>.
a. av
b. kask
c. mantar
d. çatlak

58. bottle: She found a <bottle>.
a. kap
b. bardak
c. şişe
d. küvet

59. pub: They went to the <pub>.
a. kütüphane
b. bar
c. havuz
d. barınak

60. olive: I don’t like <olives>.
a. zeytin
b. mücevher
c. çizgi film
d. mum

61. celebrate: They have <celebrated> a lot recently.
a. başarmak
b. kutlama yapmak
c. yayın yapmak
d. gelişmek

62. commemorate: We must <commemorate> him.
a. kınamak
b. tanıtmak
c. sakinleştirmek
d. anmak

63. interrogate: He <interrogated> the man.
a. tokat atmak
b. zorlamak
c. sorguya çekmek
d. rüşvet vermek

64. review: We <reviewed> the plan.
a. onaylamak
b. gözden geçirmek
c. reddetmek
d. uygulamak

65. blind: He is <blind>.
a. kör
b. ünlü
c. zeki
d. sabırlı

66. haunt: The house was <haunted>.
a. uygun
b. boşaltılmış
c. yalıtımlı
d. perili

67. shudder: The boy <shuddered>.
a. hıçkırarak ağlamak
b. burnunu çekmek
c. ürpermek
d. inlemek

68. jump: She tried to <jump>.
a. karar vermek
b. zıplamak
c. tırmanmak
d. hareket etmek

69. leaf: He touched the <leaf>.
a. yaprak
b. enkaz
c. kafatası
d. heykel

70. legend: It is now a <legend>.
a. yük
b. liman
c. efsane
d. ceza

71. lake: Many people visit this <lake>.
a. şehir
b. kilise
c. orman
d. göl

72. beneficial: These six foods are <beneficial>.
a. benzer
b. faydalı
c. yapay
d. öldürücü

73. soldier: He is a <soldier>.
a. konuk
b. bakan
c. bakkal
d. asker

74. vocabulary: They have limited <vocabulary>.
a. sözcük
b. hoşgörü
c. görev süresi
d. cephane

75. shoe: Where is your <shoe>?
a. kravat
b. çanta
c. palto
d. ayakkabı

76. mug: They gave us these <mugs>.
a. atkı
b. kulplu bardak
c. kupon
d. şeker

77. independence: Too much <independence> is bad for a child.
a. eleştiri
b. ilgi
c. özgürlük
d. övgü

78. nun: We saw a <nun>.
a. hendek
b. rahibe
c. cüce
d. çeşme

79. disguise: She <disguised> herself.
a. kılık değiştirmek
b. bıçaklamak
c. üzmek
d. çimdiklemek

80. handle: I can't <handle> it.
a. açmak
b. başa çıkmak
c. hatırlamak
d. inanmak

81. scrub: He is <scrubbing> it.
a. fermuarını kapatmak
b. kopya etmek
c. içine çekmek
d. ovmak

82. immigrate: The man <immigrated>.
a. göç etmek
b. boğulmak
c. yalvarmak
d. prova yapmak

83. congest: The roads were <congested>.
a. ücretli
b. temizlenmiş
c. kalabalık
d. aşınmış

84. devastate: They were <devastated>.
a. perişan
b. uygar
c. ayrıcalıklı
d. engelli

85. corpse: They found the <corpse> in the park.
a. giysi
b. güvercin
c. rozet
d. ceset

86. circle: Make a <circle>.
a. iddia
b. sepet
c. daire
d. zarar

87. stone: He sat on the <stone>.
a. sandalye
b. taş
c. yer
d. yatak

88. peel: Shall I <peel> it?
a. atlamak
b. hoş görmek
c. ilan etmek
d. kabuğunu soymak

89. abandon: They <abandoned> it.
a. terketmek
b. iptal etmek
c. inşa etmek
d. keşfetmek

90. current: These are our <current> prices.
a. indirimli
b. asıl
c. güncel
d. tavsiye edilen

Levels Test

07.01.2016

Ad-Soyad: ____________________

Öğrenci No: ____________________

Sınıf: ____________________

İmza: ____________________

Bu testlerin sonuçlarından elde edilecek veriler, bir yüksek lisans tez çalışmasına kaynak oluşturacaktır.

Bu bir **İngilizce sözcük bilgisi** testidir.

Lütfen **Türkçe ve koyu olarak yazılmış kelimelerin İngilizce karşılıklarını** veren seçenekleri (1, 2, 3, 4, 5 veya 6) boşluklara yazınız.

Örnek Soru

1 roar
2 change __5__ **çalmak**
3 elect __4__ **planlamak**
4 plan __2__ **değiştirmek**
5 steal
6 win

Eğer kelimenin anlamını bilmiyorsanız, lütfen o soruyu boş bırakınız ve diğer soruya geçiniz.

Eğer, sözcüğün anlamı hakkında herhangi bir tahmininiz var ise, soruya yanıt vermeye çalışınız.

Katılımınız için çok teşekkür ederim.

Teste geçebilirsiniz. BAŞARILAR…

Okt. Sezen AKSU BALAMUR

1. bath
2. chair ___ **şişe**
3. flower ___ **taş**
4. watch ___ **çiçek**
5. stone
6. bottle

1. patience
2. fever ___ **sürü**
3. leaf ___ **yaprak**
4. wreck ___ **sabır**
5. flock
6. greenhouse

1. appreciate
2. prove ___ **sürdürmek**
3. refuse ___ **uyarmak**
4. maintain ___ **reddetmek**
5. warn
6. select

1. vocabulary
2. corpse ___ **sözcük**
3. badge ___ **hayranlık**
4. ammunition ___ **ceset**
5. sorrow
6. awe

1. weave
2. label ___ **gözden geçirmek**
3. implement ___ **kutlama yapmak**
4. review ___ **dokumak**
5. evolve
6. celebrate

1. delicious
2. awful ___ **fakir**
3. poor ___ **lezzetli**
4. strong ___ **yuvarlak**
5. round
6. huge

1. basis
2. circle ___ **çekmece**
3. battery ___ **daire**
4. claim ___ **temel**
5. drawer
6. hook

1. novel
2. reward ___ **efsane**
3. legend ___ **çare**
4. harbor ___ **ödül**
5. mechanic
6. remedy

1. pub
2. library ___ **asker**
3. guest ___ **sonuç**
4. soldier ___ **bar**
5. comment
6. result

1. sob
2. shudder ___ **ürpermek**
3. rove ___ **ovmak**
4. chatter ___ **dolaşmak**
5. zip
6. scrub

1. wait
2. swim ___ **görmek**
3. play games ___ **yeri kazmak**
4. dig ___ **araba kullanmak**
5. see
6. drive

1. weep
2. boast ___ **azalmak**
3. deteriorate ___ **ağlamak**
4. diminish ___ **göç etmek**
5. immigrate
6. rehearse

1. request
2. audience ___ **davranış**
3. deficit ___ **bileşik**
4. investment ___ **hesap açığı**
5. behavior
6. compound

1. grow
2. speak ___ **başa çıkmak**
3. remember ___ **büyümek**
4. handle ___ **zıplamak**
5. jump
6. decide

1. peculiar
2. beneficial ___ **kırılabilir**
3. insulated ___ **perili**
4. haunted ___ **faydalı**
5. alike
6. fragile

1. congested
2. eroded ___ **kalabalık**
3. fragrant ___ **kuşkulu**
4. dubious ___ **çok yönlü**
5. versatile
6. beloved

1. independence
2. solution ___ **köylü**
3. peasant ___ **çözüm**
4. criticism ___ **özgürlük**
5. victory
6. refugee

1. idea
2. joke ___ **zaman**
3. bag ___ **espri**
4. shoe ___ **ayakkabı**
5. food
6. time

1. dwarf
2. nun ___ **katliam**
3. drought ___ **güneş tutulması**
4. slaughter ___ **rahibe**
5. aide
6. eclipse

1. lonesome
2. weak ___ **kör**
3. blind ___ **yalnız**
4. smart ___ **güncel**
5. current
6. reduced

1. proclaim
2. peel ___ **kılık değiştirmek**
3. envy ___ **imrenmek**
4. pinch ___ **kabuğunu soymak**
5. disguise
6. humiliate

1. furniture
2. speech ___ **basınç**
3. pressure ___ **diz**
4. cash ___ **konuşma**
5. toy
6. knee

1. flour
2. leisure ___ **mermer**
3. marble ___ **hırs**
4. ambition ___ **zeytin**
5. olive
6. jewellery

1. latter
2. loyal ___ **yetkin**
3. civilized ___ **perişan**
4. devastated ___ **ikinci**
5. competent
6. most significant

1. danger
2. pain ___ **göl**
3. lake ___ **tehlike**
4. church ___ **bitki**
5. animal
6. plant

1. bribe
2. interrogate ___ **anmak**
3. wink ___ **alkışlamak**
4. denounce ___ **sorguya çekmek**
5. applaud
6. commemorate

1. threshold
2. coffin ___ **kayış**
3. strap ___ **çatlak**
4. pillow ___ **eşik**
5. fracture
6. prey

1. knight
2. crook ___ **yengeç**
3. vinegar ___ **kulplu bardak**
4. crab ___ **sahtekar**
5. scarf
6. mug

1. abandon
2. explore ___ **iddia etmek**
3. translate ___ **taş döşemek**
4. pave ___ **terketmek**
5. capture
6. allege

1. develop
2. lend ___ **ödünç vermek**
3. seal ___ **direnmek**
4. recommend ___ **mühürlemek**
5. demand
6. resist

Yes-No Test

07.01.2016

Ad-Soyad: ____________________

Öğrenci No: ____________________

Sınıf: ____________________

İmza: ____________________

Bu testlerin sonuçlarından elde edilecek veriler, bir yüksek lisans tez çalışmasına kaynak oluşturacaktır.

Bu bir **İngilizce sözcük bilgisi** testidir.

Lütfen **verilen İngilizce kelimelerden anlamlarını bildikleriniz varsa** bunların yanındaki kutucuğa ☑ işareti koyunuz. Eğer kelimenin anlamını **bilmiyorsanız veya emin değilseniz,** herhangi bir **işaretleme yapmayınız**.

Katılımınız için çok teşekkür ederim.

Teste geçebilirsiniz. BAŞARILAR…

Okt. Sezen AKSU BALAMUR

1 □ see	11 □ maintain	21 □ crab	31 □ review
2 □ glandle	12 □ litholect	22 □ peasant	32 □ strap
3 □ time	13 □ result	23 □ acklon	33 □ galpin
4 □ threshold	14 □ humberoid	24 □ abandon	34 □ latter
5 □ connery	15 □ poor	25 □ vocabulary	35 □ congest
6 □ eclipse	16 □ bodelate	26 □ knee	36 □ shoe
7 □ deficit	17 □ fragile	27 □ dowrick	37 □ compound
8 □ adair	18 □ batcock	28 □ dig	38 □ immigrate
9 □ speech	19 □ lonesome	29 □ weep	39 □ commemorate
10 □ scrub	20 □ joke	30 □ fracture	40 □ seal

1 □ lake	11 □ cantileen	21 □ pressure	31 □ patience
2 □ pave	12 □ ambition	22 □ cambule	32 □ round
3 □ olive	13 □ jump	23 □ dubious	33 □ allege
4 □ eckett	14 □ legend	24 □ rove	34 □ shudder
5 □ drawer	15 □ devastate	25 □ stone	35 □ aistrope
6 □ remedy	16 □ pernicate	26 □ mug	36 □ lend
7 □ bastionate	17 □ warn	27 □ awe	37 □ solution
8 □ opie	18 □ escrotal	28 □ basis	38 □ jarvis
9 □ haunt	19 □ nun	29 □ scurrilize	39 □ peel
10 □ drive	20 □ diminish	30 □ blind	40 □ flower

1 □ behavior	11 □ marble	21 □ current	31 □ pub
2 □ delicious	12 □ leaf	22 □ disguise	32 □ danger
3 □ draconite	13 □ celebrate	23 □ grow	33 □ recenticle
4 □ corpse	14 □ interrogate	24 □ benevolate	34 □ refuse
5 □ slaughter	15 □ bottle	25 □ weave	35 □ applaud
6 □ soldier	16 □ horobin	26 □ flock	36 □ competent
7 □ troake	17 □ crook	27 □ scudamore	37 □ handle
8 □ versatile	18 □ contrivial	28 □ plant	38 □ beneficial
9 □ resist	19 □ envy	29 □ stimulcrate	39 □ independence
10 □ fluctual	20 □ circle	30 □ reward	40 □ nonagrate

Appendix 2: Specifications for New Items

Example Item:

school: This is a big **school**.
a. where money is kept
b. sea animal
c. place for learning
d. where people live

Overall

- The target word is presented in isolation and in bold within a context sentence
- The answer key should be randomly generated
- Avoid gender-biased language and have balanced gender representation

Target words

- Written in citation form
- From frequency list based on established corpus (BNC/COCA)
- Random sampling of words from each word-frequency level

Context sentence

- Context sentences in the first two 1,000-word levels should be written using vocabulary within the first 1,000-word level whenever possible
- Context sentences in the third 1,000-word level and above should be written using vocabulary within the first two 1,000-word levels whenever possible
- In cases where the part of speech is ambiguous, the most common form should be used based on frequency data
- The accompanying sentence should be as contextualized as possible without giving hints to the meaning of the target word

Distractors

- Core meanings of distractors should be of similar word frequency and difficulty level as the target word
- Distractors for items in the first two 1,000-word levels should be written using vocabulary within the first 1,000-word level whenever possible
- Distractors in the third 1,000-word level and above should be written using vocabulary within the first two 1,000-word levels whenever possible
- To as great a degree as possible, all distractors should be equally plausible in the context sentence

Appendix 3: Target Words

1K	2K	3K	4K	5K
bottle	basis	abandon	ambition	applaud
danger	blind	allege	beneficial	awe
delicious	lonesome	behavior	diminish	commemorate
dig	circle	celebrate	disguise	congest
drive	current	competent	envy	corpse
flower	drawer	compound	flock	crab
grow	knee	deficit	fracture	crook
handle	pressure	devastate	fragile	dubious
joke	pub	independence	haunt	eclipse
jump	refuse	latter	immigrate	interrogate
lake	warn	legend	leaf	mug
plant	resist	pave	marble	nun
poor	result	peasant	olive	rove
round	seal	remedy	patience	scrub
see	soldier	review	peel	shudder
shoe	lend	reward	strap	slaughter
stone	speech	solution	threshold	versatile
time	maintain	weave	weep	vocabulary

Appendix 4: Consent Form

Bilgi ve Kabul Formu

Ben Bursa Uludağ Üniversitesi Yabancı Diller Yüksekokulu İngilizce okutmanlarından Sezen AKSU BALAMUR. Bursa Uludağ Üniversitesi'nden Yrd. Doç. Dr. Meral ÖZTÜRK danışmanlığında bir araştırma yürütüyorum. Bu araştırma kapsamında, üç farklı sözcük bilgisi testi kullanarak, İngilizce'yi yabancı dil olarak öğrenen siz öğrencilerimizin İngilizce pasif sözcük dağarcığının ölçülmesini hedeflemekteyiz. Bu araştırmanın amacına ulaşabilmesi için siz değerli öğrencilerimizin anket çalışmalarına aktif katılımı gerekmektedir. Katılım gölüllülük esaslı ve sınırlı sayıda gerçekleşecektir.

Kimliğinizle ilgili bilgiler bu araştırma sonucu herhangi bir raporda yayınlanmayacaktır. Adınızla birlikte verdiğiniz cevaplar araştırmacı dışında kimse tarafından bilinmeyecektir.

Anket sorularına verdiğiniz cevaplar araştırmaya çok büyük katkı sağlayacaktır. Araştırmaya katılmak istiyorsanız, sayfanın altındaki ilgili yerleri doldurarak imzalayınız. Katkınız için sonsuz teşekkürler.

İngilizce Okutmanı Sezen AKSU BALAMUR
MA Programı
Bursa Uludağ Üniversitesi

Bu formdaki bilgileri okudum ve araştırmaya katılmayı kabul ediyorum.

Adım Soyadım:

Bölümüm:

Sınıfım:

İmza:

Tarih: 15.02.2016

CURRICULUM VITAE

PERSONAL INFORMATION

Place of Birth	: Uzunköprü / EDİRNE
Date of Birth	: 15.01.1980

EDUCATION

2014 - 2019	Bursa Uludag University, Institute of Educational Sciences, M.A. in English Language Education
1998 - 2002	Bursa Uludag University, Faculty of Education, Department of Foreign Languages, English Language Teaching (high honor student)
1994 - 1998	Tekirdag Anatolian Hotel Management and Tourism Vocational High School (the top student of the school)

COURSES, CONFERENCES, AND SEMINARS ATTENDED

Cambridge University Press Conference: Oracy Skills	26.04.2019
A Pearson event entitled "Next Generation Learning" by Tony Gurr	17.12.2016
Bursa Uludag University, School of Foreign Languages, 2ndInternational FLT Conference *ELT Matters*	07.05.2016
ULEAD 2015Annual Congress	8-10.10.2015
5th International Conference on Research in Education-ICRE (Presenter) at Edirne Trakya University	
Bursa Uludag University, School of Foreign Languages, 1st International FLT Conference *ELT Matters*	26.04.2014
On the Threshold of Excellence X (Seminar)	05.04.2014
Propell Workshop for the TOEFLIBT Test, Listening, Reading, Speaking and Writing	16.01.2014

Oxford ELT Conference	10-11.04.2013
The 1st International CBUTEFL Conference: “Learning, Teaching and Research in EFL”	30.10.2010
Pearson-Longman International Teacher Training Courses on Testing	16-22.07.2007
Cambridge University Press ELT Symposium	17-20.04.2007
Make Your Voice Heard (Seminar)	07.04.2007
On the Threshold of Excellence III (Seminar)	24.03.2007
Pearson-Longman International Teacher Training Courses on Testing	24-29.07.2006

WORK EXPERIENCE

22.09.2002 -	Bursa Uludag University, School of Foreign Languages, English Instructor
02.12.2015-	Bursa Uludag University, School of Foreign Languages, Member of Board
2017-	Bursa Uludag University, School of Foreign Languages, Reading Coordinator
2017-2018	Bursa Uludag University, School of Foreign Languages, Testing Office, Professional Development and Curriculum Evaluation Committee
2011-2017	Bursa Uludag University, School of Foreign Languages, Head of the Testing Department
2006-2011	Bursa Uludag University, School of Foreign Languages, a Member of the Testing Office

HONORS AND AWARDS

Sept., 2017	Prof. Dr. Recep ÇIBIK Bursa Uludag University, Director of School of Foreign Languages (2006-2017, Testing Department)

BURSA ULUDAĞ ÜNİVERSİTESİ

TEZ ÇOĞALTMA VE ELEKTRONİK YAYIMLAMA İZİN FORMU

Yazar Adı Soyadı	Sezen AKSU BALAMUR
Tez Adı	Üç Farklı Sözcük Bilgisi Testinin Kıyası
Enstitü	Eğitim Bilimleri Enstitüsü
Ana Bilim Dalı	Yabancı Diller Eğitimi Ana Bilim Dalı
Bilim Dalı	İngiliz Dili Eğitimi Bilim Dalı
Tez Türü	Yüksek Lisans Tezi
Tez Danışman(lar)ı	Doç. Dr. Levent UZUN
Çoğaltma (Fotokopi Çekim) İzni	□ Tezimden fotokopi çekilmesine izin veriyorum. □ Tezimin sadece içindekiler, özet, kaynakça ve içeriğinin % 10 bölümünün fotokopi çekilmesine izin veriyorum. ☑ Tezimden fotokopi çekilmesine izin vermiyorum.
Yayımlama İzni	□ Tezimin elektronik ortamda yayımlanmasına izin veriyorum. ☑ Tezimin elektronik ortamda yayımlanmasının ertelenmesini istiyorum. 1 yıl ☑ 2 yıl □ 3 yıl □ □ Tezimin elektronik ortamda yayımlanmasına izin vermiyorum.

Hazırlamış olduğum tezimin yukarıda belirttiğim hususlar dikkate alınarak, fikrî mülkiyet haklarım saklı kalmak üzere Uludağ Üniversitesi Kütüphane ve Dokümantasyon Daire Başkanlığı tarafından hizmete sunulmasına izin verdiğimi beyan ederim.

12.06.2019

Sezen AKSU BALAMUR