The Development and Validation of a Lemma-Based Yes/No Vocabulary Size Test

Ahmed Masrai

Abstract
Vocabulary size measures serve important functions, not only with respect to placing learners at appropriate levels on language courses but also with a view to examining the progress of learners. One of the widely reported formats suitable for these purposes is the Yes/No vocabulary test. The primary aim of this study was to introduce and provide preliminary validity evidence for a lemma-based vocabulary size test (LVST), which has been designed to measure L2 learners’ receptive vocabulary size of English. The test was administered to 219 participants. Various validity measures were operationalized to examine the performance of the test. Results show high reliability indices for the test in terms of parallel forms and internal consistency. The results also indicate that a single dimension underlies the test construct, and that the test is capable of distinguishing learners of varying proficiency levels. Concerning the item difficulty, results show approximately a linear order of difficulty across frequency levels. The test also correlated significantly with a measure of general English proficiency. The results, collectively, suggest that the LVST can usefully measure the intended construct of receptive vocabulary size, and greatly extends the range of vocabulary size measurement provided by other Yes/No tests.

Keywords
vocabulary size, test validity, language proficiency, vocabulary test, item difficulty

Introduction
The last three decades or so have witnessed an abundance of research on measuring second language (L2) vocabulary knowledge (e.g., Milton, 2009; Nation & Beglar, 2007; Read, 2000; Schmitt et al., 2001; Wesche & Paribakht, 1996), and this area has become one of the hot topics in applied linguistics. Research has shown a considerable positive correlation between vocabulary knowledge and performance in L2 skills. Specifically, receptive L2 skills (i.e., reading and listening) have been theorized and demonstrated to be greatly dependent on vocabulary knowledge (e.g., Laufer, 1992; Milton et al., 2010; Nation, 2006; Qian, 2002). Conclusions drawn from these studies generally suggest that vocabulary knowledge is a main predictor of L2 reading and listening comprehension, and that vocabulary size is used as a proxy for general proficiency in L2 (see Alderson, 2005). Having recognized the central role of vocabulary knowledge in language comprehension and the fact that the lexicons of most languages are very large, researchers have addressed the need for vocabulary size measures to assess whether L2 learners have attained a sufficiently large vocabulary to meet the demands their academic or communicative goals place on them (e.g., Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006; Schmitt et al., 2011; van Zeeland & Schmitt, 2013).

Although a number of vocabulary tests do exist, however, not many of them are time and cost-effective for use by teachers and researchers. In the L2 classroom, there are significant limits on the time and expertise available for testing, with a premium on measures that are time and resource efficient (Mochida & Harrington, 2006). In addition to the simplicity of administration, vocabulary measures need to have certain features, such as a high sampling rate, a check for over-estimation (i.e., correction for guesswork) and an easy scoring method. A promising candidate in this regard is the Yes/No test formats, such as those developed by Meara and colleagues (e.g., Meara & Buxton, 1987; Meara & Jones, 1990; Meara & Milton, 2003).

The Yes/No tests that are frequently reported in a number of research studies as measures of L2 learners’ vocabulary size have been the X-Lex (Meara & Milton, 2003), the Eurocentres Yes/No tests (Meara & Jones, 1990; Meara & Milton, 2003).

The European Centre for Modern Language Studies (2000, 2003) has published a series of Yes/No tests, the Eurocentres Yes/No tests (Meara & Jones, 1990; Meara & Milton, 2003).

The results, collectively, suggest that the LVST can usefully measure the intended construct of receptive vocabulary size, and greatly extends the range of vocabulary size measurement provided by other Yes/No tests.
construct of the X-Lex was clearly defined by its authors, the Eurocentres vocabulary size test, in addition to being dated, is problematic because of the way it was constructed and the method of how the total score is reached (Milton, 2009). The X-Lex, however, is only limited to measuring the vocabulary size of the most 5,000 frequent lemmas, which falls short of showing the ability of more proficient L2 learners. Moreover, it was developed from two different frequency lists of work by Hindmarsh (1980) and Nation (1984), which are not as representative as newer and larger corpora. Frequency lists have been updated a number of times since those lists were generated. The LexTALE (Lemhöfer & Broersma, 2012), on the other hand, although intended to measure vocabulary size of advanced level learners, it has the problem of very low sampling rate. According to Gyllstad et al. (2015), a sampling rate of 1:100 (i.e., one item to represent other 100 in a 1,000-frequency level) or lower is problematic in a vocabulary size test development. Items in the LexTALE were sampled as follows: 1 item from the first 1,000-frequency level, four items from the second 1,000, two items from the third 1,000, eight items from the fourth 1,000, four items from the fifth 1,000, eight items from the sixth 1,000, five items from the seventh 1,000, two items from the eighth 1,000, zero item from the nineth 1,000, five items from the tenth 1,000, and two items from the eleventh 1,000. The total number of items in the LexTALE are 40. With these highlighted issues in mind, there appears to be a need for a Yes/No test that can overcome the above-mentioned shortcomings.

To this end, the current study is an attempt to introduce a new lemma-based Yes/No vocabulary size test that is capable of measuring vocabulary size among proficiency groups beyond those of earlier and more widely available Yes/No tests (i.e., X-Lex; Meara & Milton, 2003) developed using updated lemmatized frequency lists. This newly introduced test aims to be efficient and practical in terms of its administration and scoring. Since the test is developed from a reasonably large sampling rate across a wide range of frequency bands, it can be used to measure vocabulary size of L2 learners of different proficiency levels in English. This would be useful for separating learners of varying abilities in a language program and setting goals for vocabulary size. A further important consideration that was taken into account when developing the current test is the use of lemma notion as the word counting unit of L2 learners’ vocabulary size. Justifications for using such an approach are discussed in more detail later in the paper.

Review of the Literature

Defining Vocabulary Knowledge

There are various types of knowledge involved in being able to use a word appropriately and effectively in a foreign language. In the same way that we should be clear about what we mean by a word in making estimates of vocabulary knowledge, we should also be clear about what we mean by knowing (Milton, 2009). While it is hard to do justice to this particular topic in the present study, the common definitions are briefly discussed. In order to define knowledge of a word, several but generally complementary frameworks have been developed (e.g., Nation, 2001; Qian, 1999; Read, 1993; Richards, 1976). In these frameworks, a distinction that researchers in vocabulary acquisition and testing find useful has often been made between two dimensions of vocabulary knowledge: breadth (or size) and depth. Breadth of vocabulary knowledge refers to the size of vocabulary or the number of words of which a learner has at least some partial knowledge (Nation, 2001). According to Read (2004, p. 211), measures of breadth of knowledge require “just a single response to each target word, and by implication, give only a superficial indication of whether the word is known or not.” Depth of vocabulary knowledge, on the other hand, relates to how well a learner knows a word (Read, 1993), such as knowledge of spelling, pronunciation, register, as well as syntactic and collocational properties (Qian, 1999).

Similar to the distinction made between breadth and depth of word knowledge is the distinction made between receptive and productive vocabulary knowledge. Receptive knowledge refers to “what is needed to understand words when they are encountered in listening and reading, and productive knowledge is the knowledge necessary to use a word in speech and writing.” (Webb, 2013, p. 1657). Receptive knowledge is commonly measured with tests of vocabulary breadth and productive knowledge is commonly measured with tests tapping vocabulary depth. The present study operationalized the term “breadth” of word knowledge.

In addition to the distinction made between the breadth and depth of vocabulary knowledge, the term “vocabulary breadth” needs further consideration. If the purpose of vocabulary testing is to identify the mastery level of knowledge (i.e., the percentage of knowledge) of words within a 1,000-word level, then “vocabulary breadth” should be referred to as “level.” If the purpose of testing, on the other hand, is to estimate a learner’s overall vocabulary size, then “vocabulary breadth” would appropriately be referred to as “size.” Introducing a vocabulary breadth distinction to include “level” and “size” is important because many researchers in the field refer to vocabulary size and level interchangeably, while they are in fact different aspects of vocabulary breadth. For example, if a learner’s vocabulary size is 4,000 lemmas or word families, it could not be said that the learner has a mastery “level” of knowledge of the 4,000 most frequent lemmas or word families. All L2 learners, except the most able ones, would have gaps in their knowledge of items in the most frequency bands (McLean & Kramer, 2015). This brief clarification of vocabulary breadth with its reference to vocabulary levels and size is crucial in
the interpretability of vocabulary tests scores. Thus, in vocabulary research a choice of definition of vocabulary knowledge is made to fulfill a particular purpose of a particular study. For the purpose of the present study, however, vocabulary size is operationalized.

Another important issue, mentioned at the beginning of this section, is what we mean by a word in making estimates of vocabulary size. For the purpose of this study, a distinction has been made between the two most widely accepted terms in the recent literature on measuring L2 vocabulary size: lemma and word family. A lemma includes a headword and its inflected, irregular, and reduced forms (i.e., n’t) that are of the same part of speech (Francis & Kučera, 1982). For example, the lemma develop includes develop, develops, developed, and developing. Word family, on the other hand, is defined as “a base word and all its derived and inflected forms that can be understood by a learner without having to learn each form separately” (Bauer & Nation, 1993, p. 253). Given that each word family has between four and six members on average (Nation, 2006), only relatively few individual items would need to be tested to estimate knowledge of a relatively large number of words. Although most vocabulary size tests have been developed employing the word family definition of a word, a number of recent research studies (e.g., Kremmel, 2016; Kremmel & Schmitt, 2016; McLean, 2018) have shown that the word family might not be the best counting unit of vocabulary size for use with L2 learners. These studies have suggested the use of lemma as a better alternative. Further discussion about the choice of lemma for the development of the test reported in this study is provided later in the paper.

Measuring L2 Receptive Vocabulary Knowledge

Several measures of L2 receptive vocabulary knowledge have been devised in the field of L2 pedagogy and testing, particularly in the last three decades (e.g., Vocabulary Levels Test (Nation, 1990); Eurocentres Vocabulary Size Test (Meara & Jones, 1990); X-Lex test (Meara & Milton, 2003); Vocabulary Size Test (Nation & Beglar, 2007)). It has been commonly reported in the literature that these vocabulary tests can be utilized to explore if a learner possesses adequate vocabulary to perform specific tasks in different L2 skills, monitor the development of a learner’s L2 receptive vocabulary knowledge, and place learners in the appropriate level according to their vocabulary size (Beglar, 2010; Milton, 2009; Nation, 2006). Formats such as the vocabulary levels test have also been claimed to assess the extent to which a certain language program accords with its objectives and design curricula and course materials appropriate for L2 learners at distinct levels (Schmitt et al., 2001). However, current investigations (e.g., Kremmel & Schmitt, 2016; Schmitt et al., 2019) generally indicated that the above-mentioned test formats are not capable of demonstrating the vocabulary knowledge required for the purpose of reading, for example.

Nonetheless, despite the criticism of these types of tests, they remain very useful, at least for providing estimates of vocabulary size, until better replacements become available. One of the measures of L2 vocabulary size that is less complicated in terms of its content, construct, administration and scoring and still provides reliable estimates of vocabulary size is the Yes/No test (Zhang et al., 2020).

The Yes/No test measures L2 receptive vocabulary size by eliciting a simple judgment as to whether a test-taker recognizes a presented word or not. Test items are commonly sampled from a range of frequency bands, with performance at respective bands used as the basis for estimating the test-taker’s receptive vocabulary size (Meara, 1996). This test format has the merits of easy construction and allowing a large number of items to be efficiently tested and scored in a relatively small amount of time (Beeckmans et al., 2001; Mochida & Harrington, 2006). Because of these qualities, the Yes/No testing technique has been used for research purposes in second language acquisition (e.g., Hermans, 2000; Milton, 2009) and as a placement test. In addition to its primary use for measuring receptive vocabulary size (e.g., Kempe & MacWhinney, 1996), the Yes/No test format has also been used in L2 sentence completion tasks (Yule et al., 1985), lexical retrieval tasks (Snellings et al., 2002) and grammatical knowledge testing (Eyckmans, 2004).

Vocabulary items in the Yes/No test normally include real words and “pseudowords.” The latter are created to be phonologically and orthographically conceivable words in the language that are used in the test design to control for guessing, if practiced by test-takers. The pseudowords approach was first introduced by Anderson and Freebody (1983) to take into consideration the probability that certain test-takers might overestimate their vocabulary size. Thus, claiming knowledge of pseudowords should lead to adjusting the test scores downwards to give a better estimate of the vocabulary size. Another approach, which is similar to that of pseudowords, for adjusting for guesswork in Yes/No tests has been recently suggested by Masrai and Milton (2018). This approach uses very infrequent real items as control to replace pseudowords to avoid the credibility issue of pseudowords creation. In Masrai and Milton’s (2018) study, prior to including the infrequent items in the Yes/No test they developed, the chance of knowing them was examined using native-speakers enrolled in a PhD program. Masrai and Milton (2018) argue that if knowledge of these words is rejected by educated native-speakers, it is very unlikely or impossible to be known by L2 learners. This approach was also proposed by Mochida and Harrington (2006) to replace pseudowords. Nonetheless, as Yes/No tests include pseudowords or control-words, there are different possible response styles to the test items. These response alternatives in the Yes/No test are discussed in some detail later in the paper.
Validation Studies of a Yes/No Test

In this section, a number of studies that have examined the validity of the Yes/No test format are reported. Meara and Buxton (1987) and Meara (1996) examined performance on the Yes/No test against multiple-choice tests used mainly by European learners of English as a second language in the UK and reported a relatively strong positive correlation (r = .7) between the two formats. Shillaw (1999) also examined performance on the Yes/No test with scores on a multiple-choice proficiency test of 201 first year Japanese university students. The study showed a significant positive correlation (.42 to .48, p < .01) between the Yes/No and multiple-choice tests. In a study by Eyckmans (2004), which attempted to directly establish the concurrent validity between the performance on the Yes/No test and the translation of the test items, only moderate correlations were observed between the Yes/No responses and the translation task (r = .3-.5). It was also noted in Eyckman’s study that the false alarm rate (responding yes to pseudowords) ranged from 20% to 25% showing a negative, but weak, correlation between responses to real and pseudowords. This negative relationship denotes a response bias toward consistently responding “yes” or “no” to both words and pseudowords (Mochida & Harrington, 2006, p. 75). Mochida and Harrington (2006) argue that this response bias is independent of the test-takers’ vocabulary knowledge, and the reliability of the Yes/No format as a test of vocabulary depends in part on adequately controlling for it when it occurs (p. 75). The study by Mochida and Harrington (2006), which compared the performance of the Yes/No test versus that of the vocabulary levels test (VLT), concluded that the Yes/No test is a valid measure of L2 vocabulary size, with implications for L2 classroom application.

Other validation studies on the performance of the Yes/No test were also conducted by Harrington and Carey (2009), Milton et al. (2010) and Roche and Harrington (2013). Harrington and Carey (2009) evaluated the concurrent validity of an on-line Yes/No test with a battery of placement tests, including grammar, writing, speaking and listening, at an Australian English language school. They concluded that the placement level decision strongly correlated with scores from the Yes/No test; correlations ranged from .6 to .8, according to the placement test parts. Milton et al. (2010) examined the relationship between scores on a written Yes/No test (X-Lex; Meara & Milton, 2003) and performance on the International English Language Testing System (IELTS). Their results indicated strong correlations between scores on the X-Lex and the skills of reading and writing (r = .70 and .76, respectively) in IELTS. Finally, Roche and Harrington (2013) explored the prediction of a timed Yes/No test format, as a measure of vocabulary size, of written academic English proficiency and overall academic achievement among L2 learners. The study concluded that the Yes/No test is a reliable and cost and time effective measure of academic English proficiency, and can be used as an indicator of vocabulary knowledge that L2 learners need to undertake study in a tertiary level English medium program.

While using the correlational analyses approach to establish the validity of the Yes/No tests and other tests, and also associating scores on these tests to general language proficiency has been utilized for over two decades or so, correlations alone are not sufficient to confirm a test validity. In fact, it is not expected to always find strong correlations between different test formats, and this claim is supported by the wide range of correlations observed in the discussion above. Furthermore, the correlations between vocabulary size and general language proficiency, for example those found in Milton et al. (2010) between X-Lex and IELTS, would not qualify a vocabulary size test or a cloze test to be an accurate proxy of language proficiency. Vocabulary size only assesses one aspect of the whole language performance, albeit very important one.

The present study, however, operationalized various aspects of validity measures, such as those described by Messick (1989, 1995), for the newly developed test. Messick’s validity framework has been recently utilized by a number of studies to examine the validity of vocabulary size tests (e.g., Beglar, 2010; Zhao & Ji, 2018). This framework includes the substantive aspect of construct validity, the structural aspect of construct validity and the assessment of differences in the performance of groups of distinct abilities. The substantive aspect of construct validity involves the “theoretical rationales for observed consistencies in test responses . . . along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks” (Messick, 1995, p. 745). The structural aspect of construct validity, on the other hand, concerns the assumed dimensionality of the construct being measured (Messick, 1989). Finally, the assessment of differences in the performance of groups of distinct abilities (Messick, 1995) concerns the capability of a test to effectively distinguish between test-takers of different of proficiency levels—that is, separate learners according to their vocabulary size.

Correction for Guessing Techniques

The inclusion of pseudowords in a Yes/No test format has implications for calculating the test score. Therefore, various techniques have been developed accordingly. Before discussing these techniques, the possible test-takers’ responses to the test items are first provided:

- Hit: responding yes to a real word;
- False alarm: responding yes to a pseudoword;
- Miss: responding no to a real word;
- Correct rejection: responding no to a pseudoword.

Among these responses, false alarms are considered crucial to the scoring and interpretation of the Yes/No test
performance. High false alarm rates indicate a high level of guessing and low reliability of a test’s scores, while very low rates of false alarms indicate an overly cautious performance that yields an underestimated vocabulary size of a test-taker (Huibregtse et al., 2002). For the case of less extreme performance, a given false alarm rate can yield a different correction effect, depending on the relative hit rate and the type of correction technique applied (Beeckmans et al., 2001; Eyckmans, 2004; Huibregtse et al., 2002).

The simplest approach to score performance in the Yes/No test is by subtracting the proportion of false alarms from the proportion of hits (i.e., h–f). However, according to Zimmerman et al. (1977), this scoring formula is too simplistic. Thus, they proposed the use of “signal detection theory” to account for the probability of response errors when accomplishing a yes/no task. This theory seeks to quantify the ability to differentiate between an important, real signal and noise, that is any error or undesired random disturbance of a useful information signal (Pellicer-Sánchez & Schmitt, 2012, p. 3). Zimmerman et al.’s study suggested that the “signal detection” technique was promising for accounting for response errors. Meara (1989), Meara and Buxton (1987) and Cameron (2002) used a correction for guessing (cfg) formula, which takes into consideration individual differences in the proportion rates of hits and false alarms. The cfg formula (see equation 1) is the proportion of h–f divided by the proportion of correctly rejected pseudowords. The obvious drawback of the cfg formula is that it stresses the hit rate over the false alarm rate, where at the most extreme, when the hit rate equals 1, the adjusted score will be 1 regardless of the proportion of false alarms (Mochida & Harrington, 2006).

1. Correction for guessing (cfg) formula:

\[ P(h) = \frac{(h) - (f)}{1 - (f)} \]

Where \( P(h) \) = true hit rate, \( h \) = observed hit rate, \( f \) = observed false alarm rate.

Taking into account the limitation of the cfg formula, Meara (1992) proposed an alternative technique for scoring the Yes/No test, that is \( \Delta m \) (see equation (2)). This formula proved to be overly conservative, producing scores that overcorrected for false alarms and uninterpretable when hits are relatively low and the false alarms are relatively high (Huibregtse et al., 2002, p. 245).

2. \( \Delta m \) formula:

\[ \Delta m = \frac{(h - f)}{(1 - f)} \cdot \frac{f}{h} \]

Since both the cfg and the \( \Delta m \) formulas overlooked the individual response bias as a source of variability, Huibregtse et al. (2002) proposed a more complex formula (see equation (3)), that is “index of signal detection” (ISDT) to account for guessing and response bias. Huibregtse et al. (2002) compared the three techniques proposed earlier and the ISDT using the results of a Yes/No test developed by Meara (1992) to find out which approach yielded the best estimate. Their findings revealed that the \( \Delta m \) technique always provided an underestimation of the intended standard, and the cfg technique yielded an overestimation for high hit rates. Interestingly, the simplistic h–f (Zimmerman et al., 1977) method produced scores that were in most cases comparable and often identical to those of the sophisticated ISDT method.

3. ISDT formula:

\[ ISDT = 1 - \frac{4h(1 - f) - 2(h - f)(1 + h - f)}{4h(1 - f) - (h - f)(1 + h - f)} \]

Although valid theoretical grounds are available for favoring one scoring technique over another (for more detail see Huibregtse et al., 2002), the real differences in application of these provided approaches are fairly small, with scores produced using \( \Delta m \) being generally low and those using cfg being generally high (Mochida & Harrington, 2006). The general conclusion of the studies comparing the aforementioned approaches for scoring the Yes/No test format is that there is no consensus on the best scoring system; however, recent researchers (e.g., Harrington & Carey, 2009; Roche & Harrington, 2013) have adopted the less complicated h–f method. Thus, based on the discussions of different approaches used to score the Yes/No test format, the current study, similar to the recent ones (e.g., Harrington & Carey, 2009; Roche & Harrington, 2013), adopted the h–f method to score the newly developed vocabulary size test. This scoring system is less complex, which is particularly effective for use within the limits of the time and expertise available for testing in language classroom.

The New Lemma-Based Vocabulary Size Test

The primary purpose of the current study is to introduce and provide preliminary validity evidence for a new lemma-based Yes/No English vocabulary size test, the LVST. The test is intended as a measure of L2 learners’ vocabulary size of English lexis from the first twelve 1,000-word frequency bands of the Corpus of Contemporary American English (COCA). The 1,000 to 12,000 word-frequency bands are measured using 20 real items and five control items per band. The 20 items in each frequency band were principally sampled. For example, the first appearing item in the band is more frequent than the second appearing item, the second is more frequent than the third, and so on. Also, one item was selected from every 50 word-group in ascending order to avoid clustering of items in a particular part of each
1,000-word frequency band. Furthermore, as the list from COCA provides information about each word part-of-speech, this was taken into account in the items sampling process to avoid bias toward a specific word class. This sampling procedure would allow generating a number of parallel versions of the same test. The test includes 300 items in total, 240 real words, and 60 control words. The 300-item test can be completed in 10 to 15 minutes. Because the test measures words at high frequency bands and fairly low frequency bands, L2 learners of varying abilities can take the test. However, whether a shortened version of the test could only be used with lower proficiency level learners is subject to further empirical evidence.

The LVST was developed using the lemma definition of counting words. The reason behind using the lemma in the development of the current test, and not the word family, is that the lemma is argued to be the best general unit of estimating vocabulary size. Schmitt (2010) offers a number of reasons for the need to use the lemma in vocabulary size testing. First, what forms are included within a lemma is straightforward relative to the word family, enabling the easy understanding of research studies. Second, the comparatively straightforward lemma makes the replication and comparison of studies more feasible. Third, the lemma is a pragmatic compromise when estimating both receptive and productive studies. Lastly, Schmitt (2010) argues that, relative to the word family, the lemma offers more realistic estimates of the large number of words required to function in English.

The notion of “lemma,” mentioned above, is supported by some recent research (e.g., Kremmel, 2016; Treffers-Daller et al., 2016), which indicates that the lemma might provide a sensible balance between grouping words together to some degree while at the same time offering a higher degree of confidence that the word on the vocabulary test does truly represent knowledge of the other words in that group. Kremmel (2016, p. 978) argues that “because the members of a lemma differ only in grammatical form rather than lexical-semantic properties (at least in some cases), knowledge of the lemma representative on the test would most likely imply knowledge of the other lemma members.” Pinchbeck (2016) appears to support this argument by suggesting that lemmas might be the most practical unit for most general test purposes. In a relatively recent study, Laufer and Cobb (2020) examined the knowledge of derived words needed for reading. Their results show that reaching the lexical thresholds for reading does not necessarily require the mastery of most of the derived words in a word family. The findings indicated that the necessary coverage for reading can be achieved with the knowledge of base-words and their inflections plus only a small number of frequent affixes. Additionally, in the measures of lexical diversity, Treffers-Daller et al. (2016) suggest that lemmas hold advantage over and above word families in disambiguating student proficiency levels. Aitchison (2012) also offers evidence for the psycholinguistic reality of lemmas in first language (L1) speakers, which is probably the case in L2 learners. With L2 learners, it is too optimistic for us to admit that if a learner knows the headwords, s/he would know all the other members of the same word. Therefore, the concept of lemma is thought to provide a strong basis for developing vocabulary size tests for use with L2 learners and offers more accurate estimates and enhanced interpretability of test scores than the word family notion.

The Source of Target Vocabulary

The items included in the LVST were sampled from lemmatized frequency lists drawn from the COCA (Davies, 2008). The COCA, however, is one of the largest up-to-date available corpora of English and is a balanced corpus of American English. This corpus is continuously growing in size, and it is reported on the corpus website that it currently contains 560 million words. Most importantly is that the frequency lists from the COCA are based on a systematic collection of texts in English from a variety of genres, including spoken language from the most prominent variation of English globally, and are lemmatized (Davies, 2008).

The LVST covers the first twelve 1,000-word frequency bands of the COCA word lists, which are expected to provide an adequate coverage for various reading genres. McLean (2018) argues that in cases where the word family provides 98% coverage of texts, the lemma provides less coverage of the same texts. Thus, if the 98% text coverage is obtained with a vocabulary size of 8,000 to 9,000-word families (Nation, 2006), it is argued that greater lemmas are required to provide such coverage. Rigorous empirical evidence, however, is still needed to inform our understanding of the relationship between lemma size, text coverage and the level of text comprehension. Nonetheless, assessing vocabulary size at the 12 most frequent bands of lemmas is thought to represent the greatest range in vocabulary learning for the vast majority of L2 learners and tell us about the extent of vocabulary size that most able learners have.

To this end, the intention of the current paper is to provide preliminary validity evidence for a test designed to measure the receptive vocabulary size, the LVST. This task will be approached by providing answers to the following research questions:

1. Can acceptable levels of reliability in the test-takers’ performance on one form and parallel forms of the test be obtained?
2. Does a single primary construct, vocabulary size, underlie the test?
3. Does the test effectively distinguish learners of different proficiency levels?
4. Is there any meaningful difficulty order among the frequency levels?
5. Can the test predict performance in a general language proficiency test?
Method

Participants

Four groups of participants (N=244) took part in the current study. The first group included 25 native speakers of English studying in doctoral programs at a British university, who were used for the purpose of piloting the test. The second, third and fourth groups included 219 non-native speakers of English, with L1 Arabic background, enrolled in a range of English language studies programs. The participants were pursuing three different degree levels. Of these, 48 students were pursuing their doctoral studies, 76 were pursuing their master’s studies, and 95 were enrolled in bachelor’s degree programs. Based on their degree programs, the non-native participants were assigned to the three groups. Group one, the high-proficiency group, included the doctoral students; group two, the mid-proficiency group, included the master’s students; and finally, group three, the low-proficiency group, included the bachelor’s degree students. These participants’ levels of proficiency were determined based on their level of study. It is worth noting here that there is a probability that participants will not always conform to group membership in terms of the level of study. In the case of the participants in this study, there were only a few who did not fall within this classification. However, to confirm the suitability of degree level for assigning the participants as a group into the proficiency levels, their performance on the Oxford Quick Placement test (OQPT) was taken into consideration. The result of a one-way ANOVA confirmed that the difference in groups’ performance on the OQPT was statistically significant, F(2, 216)=85.86, p < .001.

Instruments

English vocabulary size test. Two parallel versions of the LVST, A and B, were developed for the purpose of the study (see Appendices A & B). Each version of the test includes 240 words and 60 control items used to adjust for guessing. The items were assigned equally, based on their frequency, into 12 columns—each represents a 1,000-word band—with 20 words and 5 control items in each column. To reiterate, the items were arranged from the most frequent lemma in band 1 to the lowest in band 12. The control items were very rare, beyond the 25,000-word level in Thorndike and Lorge’s (1944) word list, and were selected from Goulden et al.’s (1990) test. Results from both Goulden et al.’s (1990) and Milton and Treffers-Daller’s (2013) studies indicate that knowledge of these words among native university populations is negligible and learners would not know them. Prior to including the control items in the test, they were piloted with a number of native speakers (N=25, doctoral students) to examine the probability of knowing them. The potential of knowing the control items included in the test was rejected by those students.

Taking the discussion of the various methods used in scoring the Yes/No test into consideration, this study followed the h-f scoring system. It is worth noting here that “miss” (responding “no” to a real word) and “correct rejection” (responding “no” to a control item) do not contribute to the total adjusted score. As the test is developed to estimate a test-takers’ vocabulary size out of 12,000 lemmas, each yes response to a real word (240 words in total) is given a credit of 50 points, that is, knowledge of one word stands for knowledge of 50 other words, and each false alarm, responding yes to a control item, is given 200 points (60 items in total). Therefore, to calculate the final adjusted score of the LVST, the proportion of false alarms (the sum of the multiplication of yes responses to control items by 200) is subtracted from the proportion of hits (the sum of the multiplication of yes responses to real words by 50). The total adjusted score is intended to provide a vocabulary size estimate of the test-taker. The maximum possible score on the LVST is 12,000.

General language proficiency test. The participants took the Oxford Quick Placement test (OQPT), which is used by some institutions to assign students to English course levels or as a requirement for admission in academic programs delivered in English. It is also used in research to assess general language proficiency (e.g., Lemhöfer & Broersma, 2012; Wang & Treffers-Daller, 2017). It should be noted that, however, with OQPT being mainly a placement test, it cannot be emphasized that this test assesses language proficiency with pronounced accuracy, but “it should certainly provide an approximate estimate for proficiency” (Lemhöfer & Broersma, 2012, p. 328). However, OQPT provides information about students’ general language ability in relation to the Common European Framework of Reference for Languages (CEFR). The OQPT is a written test that includes 60 items. For items 1 to 5, test-takers are required to understand notices and decide where they can see them. For items 6 to 20 and 41 to 50, five short passages with blanks in the texts are provided and test-takers are required to choose the word or phrase that best fits each blank from the given answer options. For items 21 to 40 and 51 to 60, a single sentence with a blank is given for each item and a test-taker should choose the word or phrase that best completes the sentence. The OQPT is designed to measure L2 learners’ English vocabulary knowledge in various ways, such as word meanings, synonyms and antonyms, collocations, and phrases. It also measures L2 learners’ grammar knowledge, such as tense, passive voice and counterfactual knowledge. To complete the test, test-takers need about 30 to 45 minutes. The maximum possible score a learner can obtain is 60.

Procedure

Both the two versions of the LVST and the OQPT were taken by the participants under controlled conditions. As the
participants were from different institutions, arrangements were made with those who volunteered to take part in the study and the targeted institutions prior to the data collection date agreed upon for each institution. The participants from each institution were gathered in a large lecture theater. The tests were then administered after giving information about the purposes of the measures and clear instructions on how to perform the tasks. The data collection sessions of both the LVST and the OQPT lasted for about 60 minutes.

### Results

#### Results for the Test Performance

The first step toward examining the performance of the LVST was the analysis of the pilot data obtained from the 25 native English speakers prior to the test administration to non-native participants. The purpose of this data was to examine the performance of the test items in terms of responses to real and control items in the two-parallel forms. In other words, data from native speakers should help identify whether the two versions of the test are identical enough and the control items are rare enough to be known for non-native speakers. The descriptive statistics for the native speakers’ scores are displayed in Table 1.

The results show that the native speakers performed similarly in versions A and B of the test ($M=11,944$ and $11,933$, respectively) and appear to reach the maximum possible score, which is not surprising since all the participants were doctoral students. These results generally indicate that the two versions of test are almost identical. It is clear that the test was too easy for native-speaking students and, thus, would arguably be feasible for advanced non-native speakers who aspire to native-like competence. The native speakers’ responses to the control items also revealed that only one participant identified one control word as known (analogon) in both versions of the test. This can confirm that any “yes” response to the control words included in the test by non-native speakers is a false alarm since they were rejected by educated native speakers.

#### RQ1: Can Acceptable Levels of Reliability in the Test-Takers’ Performance on One Form and Parallel Forms of the Test be Obtained?

To examine the initial reliability of the test, the performance of the non-native test-takers on the parallel forms of the test was first analyzed using descriptive statistics. The results are summarized in Table 2.

|       | N  | Min | Max | M   | SD  |
|-------|----|-----|-----|-----|-----|
| Version-A | 219 | 3350 | 10,400 | 6546.12 | 1820.03 |
| Version-B | 219 | 3100 | 10,650 | 6469.41 | 1934.10 |

The mean scores of the test-takers were 6,546.12 and 6,469.41, respectively, with standard deviation of 1,820.03 and 1,934.10. These results show that the performance of the test-takers is very similar in both versions of the test. There also appears no ceiling or floor effects in the two versions of the test. To examine the difference between means, a paired-sample $t$-test was conducted. The $t$-test indicates that the difference between the mean scores was not statistically significant ($t[218]=1.65; p=.10$). Also, a strong positive correlation was found between the two versions ($r=.94$), showing a close resemblance between the two.

Further reliability measures were employed to examine the performance of the test. First, Cronbach’s alpha analysis for versions A and B of the test was run for the adjusted scores. The alpha coefficient for the two versions was .96, suggesting a relatively high reliability. As the level of reliability of the parallel forms was high, a further reliability analysis was performed for version A of the test (i.e., internal consistency). The coefficient reliability score of all items in test A was .97. Also, when Guttman Split-half was used, a reliability score of .92 was found indicating that the test has a high level of internal consistency.

#### RQ2: Does a Single Primary Construct, Vocabulary Size, Underlie the Test?

In order to evaluate the dimensionality of the test, the structural aspect of construct validity, exploratory factor analysis was carried out. To perform the analysis, the overall scores obtained from the 12 frequency-levels were submitted to factor analysis. However, prior to the analysis the appropriateness of the data for factor analysis was examined. In this
regard, the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy was 0.91, above the commonly recommended value of 0.6 (Pallant, 2007). Additionally, the Bartlett’s Test of Sphericity was statistically significant (\( p < .001 \)). These indices support the factor analysis for the data set.

A Principal Axis Factoring (PAF) method was performed and signified that only one factor could be extracted. This factor per se accounted for approximately 9.5 eigenvalues, amounting to over 79% of the total observed variance. An inspection of the scree plot of the factors (see Figure 1) showed that there is only one dominant factor and that there is a large break between the first and the second factors. The factor loadings reported in Table 3, on the other hand, clearly show that all the 12 frequency-levels yielded high loadings on the single dominant factor. It is evident from Table 3 that all factor loadings were above 0.7. Brown (2006) suggests that 0.3 is regarded as the minimum required loading. In sum, these results provide evidence that only a single factor (i.e., vocabulary size) is underlying the LVST construct and that loadings are uniformly high on this factor.

**RQ3: Does the Test Effectively Distinguish Learners of Different Proficiency Levels?**

In order to answer this question, the mean scores of the three groups on the LVST were compared. Descriptive statistics for the groups’ performance are shown in Table 4. The mean scores of the high-, mid- and low-proficiency groups were 9,091.66, 7,019.74, and 4,881.05, respectively. These vocabulary size scores indicate that the proficiency of the participants in this study nearly range from intermediate to very advanced levels. An examination of the results indicates that there is a noticeable difference in the performances of the three proficiency groups. More specifically, the high-proficiency group performed better than the mid-proficiency group, and the mid-proficiency group performed better than the low-proficiency group.

To assess the statistical significance of the differences in means of the three groups, a one-way analysis of variance (ANOVA) was used. The ANOVA results revealed that the difference between the groups was statistically significant (\( F[2, 216] = 488.86, p < .001 \)). The eta squared effect size was .82, which, according to Cohen’s (1988) guidelines, denotes a large effect. Post hoc analyses, using both the Scheffe test and the more robust Games-Howell test (Field, 2009), indicated that all mean differences between the groups were statistically significant (see Table 5). These results suggest that the LVST can effectively distinguish test-takers of different proficiency levels. Further evidence to support these results is provided by the groups’ performance on the OQPT (see Tables 4 and 6).

**RQ4: Is There Any Meaningful Difficulty Order Among the Frequency Levels?**

One aspect of construct validity, the substantive aspect, concerns the rationale for the item difficulty and the person ability. In the development of vocabulary size tests that use word frequency as the base for item selection, it is generally assumed that test-takers would perform better on the higher frequency levels than the lower frequency levels (e.g., Beglar, 2010; Milton, 2009). The proposed order of difficulty in the LVST was the first 1,000 < 2,000 < 3,000 < 4,000 and so on to the 12th band. The results displayed in Table 7 (graphically illustrated in Figure 2) indicate that the mean scores of the number of known items (\( max = 20 \)) on the first to the twelfth frequency levels descended from 18.65 to 2.75, showing an order of difficulty across the frequency levels, although not very linear in the lower frequency levels. It is evident from these results that the test-takers had better performance on the most frequent words than the less frequent ones, which lends support to the idea that the vocabulary learning process occurs according to word frequency.

To examine if the differences in performance in each frequency level had any statistical significance, a one-way repeated-measures ANOVA was conducted. The results of Greenhouse–Geisser indicated that there was a significant difference between the mean performance of the test-takers on different difficulty levels (\( F[3.5, 763.78] = 1,995, p < .001, \epsilon = .90 \)). To examine those differences in more detail, pairwise comparisons were performed. The results showed that all comparisons were found to be significantly different (\( p < .001 \)).

In sum, these results provide ample evidence that the LVST appears capable of predicting learners’ performance in relation to word frequency levels and that there is a meaningful order of difficulty among these levels.
RQ5: Can the Test Predict Performance in a General Language Proficiency Test?

The sensitivity of the LVST as an approximate predictor of general language proficiency was assessed first by examining its correlation with the participants' performance on the OQPT and then by performing regression analysis. The participants' scores on both measures are summarized in Table 8. The results revealed that scores on the LVST have a significant positive correlation with performance on the general language proficiency test, as measured by OQPT, \((r = .65, p < .001)\). The regression analysis also showed that scores on the LVST can explain 42% of the variance in general language proficiency. Although this predictive value appears considerable, there are certainly other components, over and above vocabulary size, which account for general language proficiency. Statistically speaking, however, the correlation and regression results generally suggest the efficacy of the LVST as a rough indicator of performance on the OQPT, lending support to earlier findings (e.g., Lemhöfer & Broersma, 2012; Milton et al., 2010; Roche & Harrington, 2013) that vocabulary size estimates of Yes/No tests can usefully predict performance on general language proficiency tests. Yet it can be claimed that a Yes/No test, or another vocabulary size test, is a measure of general language proficiency.

Discussion and Conclusions

The aim of this article was to introduce and provide preliminary validity evidence for a new Yes/No English vocabulary size test, the LVST. The test is primarily designed to measure written receptive vocabulary size of the first 12,000 lemmas of English. The use of lemma notion in the development of the LVST is in line with the recent argument that lemma could be a useful counting unit for item sampling in L2 vocabulary size tests (e.g., Kremmel, 2016; Kremmel & Schmitt, 2016). The usefulness of the test as an approximate predictor of general language proficiency was also examined in the current study.
In order to assess the initial reliability and validity of the test, a number of measures were employed. Messick (1989) pointed out that a validity argument of a test should include both theoretical rationales and an empirical analysis. On the theoretical ground, the LVST items appear to be effectively sampled from a relatively large corpus, the COCA, and the difficulty order of the items is indexed by their frequency in the generated lists. The test also features a high level of practicality in terms of both administration and scoring (e.g., Mochida & Harrington, 2006). Little technical training is required for the administration of the test and the scoring procedure is simple. Also, the time required to complete the test is feasible, about 10 to 15 minutes. Thus, it can be concluded that the LVST is practical enough.

On the empirical ground, various analyses were operationalized. As this study presents the initial validation of a newly designed test, two parallel versions of the test were developed, forms A and B. Descriptive statistics and t-test were used to examine the difference in scores obtained using both forms. The results indicated that both forms are almost identical and that the difference between means was not statistically significant. Cronbach’s alpha coefficient suggested high reliability indices between the two forms and the split-halves of version A of the test ($\alpha = .96$ and .92, respectively). The reliability is deemed to be a fundamental requirement of a validity argument as suggested by both the Medical Outcomes Trust Scientific Advisory Committee (1995) and Messick (1995). The results also showed that no floor or ceiling effects were present, suggesting that the test can be used to estimate the vocabulary size of learners with a wide range of proficiency levels.

The preliminary validity of the LVST was operationalized using various aspects of construct validity described by

### Table 6. Post hoc Tests: Multiple Comparisons Between the Groups (OQPT).

| (I) Level (J) Level | Mean Difference | Std. Error | Sig. | 95% Confidence Interval |
|---------------------|-----------------|------------|------|-------------------------|
| Scheffe             |                 |            |      |                         |
| 1 2                 | $-7.72^{*}$     | 0.96       | 0.000| $-10.08$ $-5.36$       |
| 3 1                 | $-13.92^{*}$    | 1.10       | 0.000| $-16.64$ $-11.21$      |
| 2 3                 | $-6.20^{*}$     | 1.15       | 0.000| $-9.03$ $-3.38$        |
| 3 2                 | $13.92^{*}$     | 1.10       | 0.000| $11.21$ $16.64$        |
| Games-Howell        |                 |            |      |                         |
| 1 2                 | $-7.72^{*}$     | 0.95       | 0.000| $-9.96$ $-5.48$        |
| 3 1                 | $-13.92^{*}$    | 1.16       | 0.000| $-16.71$ $-11.14$      |
| 2 3                 | $-6.20^{*}$     | 1.34       | 0.000| $-9.39$ $-3.02$        |
| 3 2                 | $13.92^{*}$     | 1.16       | 0.000| $11.14$ $16.71$        |

Note. *The mean difference is significant at the 0.05 level.

### Table 7. Item Difficulty and Performance Across Frequency Levels.

| Frequency level | Min | Max | $M$  | SD  |
|-----------------|-----|-----|------|-----|
| First           | 14  | 20  | 18.65| 1.44|
| Second          | 11  | 20  | 17.95| 1.80|
| Third           | 10  | 20  | 16.75| 2.27|
| Fourth          | 8   | 20  | 15.30| 3.55|
| Fifth           | 6   | 20  | 14.21| 3.01|
| Sixth           | 4   | 20  | 12.86| 3.75|
| Seventh         | 0   | 19  | 9.33 | 4.17|
| Eighth          | 0   | 20  | 8.91 | 4.45|
| Ninth           | 0   | 17  | 7.18 | 4.19|
| Tenth           | 0   | 17  | 5.52 | 4.69|
| Eleventh        | 0   | 12  | 4.73 | 3.78|
| Twelfth         | 0   | 9   | 2.75 | 2.74|

**Figure 2.** Test-takers’ performance on each 1,000-frequency level.
Messick (1989, 1995). First, the structural aspect of construct validity, which concerns the assumed dimensionality of the construct being measured, was examined. The results of the dimensionality analysis, using PAF, revealed that a single construct underlies the test accounting for over 79% of the total variance. All the frequency levels had a high loading on this factor. Since the test items are focused on estimating the vocabulary size, it may be concluded that the single underlying construct is the vocabulary size. Another important aspect of construct validity, as suggested by Messick (1995), is the assessment of differences in the performance of groups of distinct abilities. The results of a one-way ANOVA indicated that the LVST can effectively discriminate between test-takers of different proficiency levels, that is, the high-group outperformed the mid-group, and the mid-group outperformed the low-group. In the light of this finding, this aspect of validity is also supported.

A third aspect of construct validity, the substantive aspect, was also considered in the study. As pointed out by Messick (1995), the substantive aspect of construct validity concerns the item difficulty and person ability. This aspect was explored by the analysis of whether the assumed order of item difficulty and person ability is realized in the data. The result of a one-way repeated measures indicated that the assumed difficulty order is met and a significant difference in performance was found between the frequency levels. Hence, it is evident that the substantive aspect of construct validity is also supported.

Performance on the LVST was also assessed as an approximate predictor of scores on the OQPT, a standard test of general language proficiency. The findings indicated a significant positive correlation ($r = .65, p < .001$) between performance on the vocabulary size test and the general language proficiency test, and that vocabulary size can explain about 42% of the variance in general language proficiency. These findings confirm the long-acknowledged idea that vocabulary size is a good predictor of general language proficiency in a foreign language (e.g., Laufer, 1992; Lemhöfer & Broersma, 2012; Milton et al., 2010; Qian, 1999). For example, the result of the current study is more in line with Lemhöfer and Broersma’s (2012) finding of a significant correlation ($r = .63$) between a Yes/No vocabulary size test (LexTALE) and general language proficiency as measured with the OQPT. A very similar correlation ($r = .65$) was also found between the OQPT and the LVST in the current study. The results also support that of Milton et al. (2010) who found a strong correlation ($r = .68$) between scores on the X-Lex and IELTS overall performance. The strong correlation between the LVST and the OQPT and its predictive value of the performance on the OQPT suggest, at least statistically, that the LVST could also be used to roughly indicate a learner’s general language proficiency. However, further solid evidence from empirical research is required to either confirm or reject this.

Another important matter concerns interpretability of the test scores—that is the degree to which quantitative estimates can be associated with qualitative meaning. First, it should be noted that a Yes/No test, similar to other vocabulary tests such as the VST (Nation & Beglar, 2007) and VLT (Schmitt et al., 2001), measures relatively shallow knowledge of a written word. There is much more to knowing a word than merely recognizing its form or form-meaning link (Schmitt, 2010). Therefore, test-takers’ responses on the LVST provide only a rough indication of how well a learner can read. Second, although the LVST covers a wide range of frequency levels in an ascending order, it is only intended to provide an estimate of the total vocabulary size but not scores for the individual levels in the test.

The most straightforward way to interpret the LVST scores is to view each real item as representing 50 lemmas (assuming that 20 items are used to represent each 1,000-frequency level) and each control item would counterbalance four lemmas, at a ratio of 1:4 (∼200). A test-takers’ adjusted score is reached through subtracting the raw score of false alarms from the total raw score of yes responses to real items. For example, if a test-taker recognized 50 items and responded yes to one control item, his/her adjusted score would be 2,300 ([50 × 50]–200). This adjusted score is believed to represent a test-takers’ total receptive vocabulary size. With an increasing number of studies advocating the use of lemma as the appropriate unit for estimating vocabulary size, the LVST is thought useful for this purpose, featuring a reasonably high sampling rate of 12 frequency levels from a relatively large corpus.

To conclude, the current study provides ample evidence to support the preliminary validity of the LVST, a test that assesses the written receptive vocabulary size of the first twelve 1,000-word frequency bands of the COCA. Because the test targets words at lower frequency bands, it is appropriate for use with more proficient L2 learners of English. The present study offered initial validation for the LVST with L2 learners from L1 Arabic background. Therefore, there is a need for further validation studies of the test with L2 groups from different L1 background. Additionally, further validation approaches, beyond those implemented in the present study could be conducted in a future research. These

### Table 8. Scores on the LVST and OQPT.

|       | N     | Min  | Max  | M    | SD     |
|-------|-------|------|------|------|--------|
| LVST  | 219   | 3,350| 10,400| 6,546.12 | 1,820.03 |
| OQPT  | 219   | 19   | 55   | 36.29 | 8.29   |
may include comparing scores from the LVST with scores from other Yes/No vocabulary size test, which has similar features, and also examination of this test against other test formats, such as multiple choice-based tests. Another point that maybe considered in a future research is the adjustment for scores overestimation. Only the non-word approach was implemented in the present study. The main issue with this approach is the chance that test-takers strategically identify non-words and avoid hitting them, which would lead to inflating their scores. It is suggested that further research consider multiple correction techniques for guessing, such as reaction time in combination with $h$–$f$ formulas. Furthermore, as the results show acceptable levels of validity evidence for this new test, future work may consider developing aural forms of the LVST designed to measure aural receptive vocabulary size, which would allow language teachers to directly compare learners’ vocabularies in both the written and aural modes. The LVST fills a gap in the field of L2 vocabulary assessment by providing a Yes/No measure that is capable of assessing vocabulary size from high to fairly low frequency words using the notion of lemma as a word counting unit. The evidence from the study suggests that the test is useful for estimating learners’ vocabulary size at intermediate to relatively advanced proficiency levels. The test is also thought useful for teachers in the language classroom as a tool for measuring vocabulary development, as it is time-cost effective and requires little expertise to utilize.

### Appendix A

**Lemma-Based Vocabulary Size Test (LVST)— Version A**

This test is designed to measure your written receptive vocabulary knowledge of English words

Please look at the words in the table. Some of these words are real English words and some are used as control words. Be aware that Yes responses to the control words are penalized. Please tick (✓) the words that you know or can use.

Here is an example. book ✓

*Note: Your overall score will be affected severely by ticking control words.*

| Level 1 | Level 2 | Level 3 | Level 4 | Level 5 | Level 6 |
|---------|---------|---------|---------|---------|---------|
| 1k-word band | 2k-word band | 3k-word band | 4k-word band | 5k-word band | 6k-word band |
| his | machine | secretary | meter | shared | scrutiny |
| other | modern | proper | myth | quantity | inspect |
| school | fast | payment | motivation | glory | reasonably |
| leave | sky | recover | appoint | seal | mature |
| kef | corium | palar | agoura | smetic | banausic |
| play | variety | opponent | grain | bull | passionante |
| kind | stone | foundation | frequent | boil | affordable |
| together | contract | illness | painter | accomplishment | persistent |
| war | vehicle | severe | decrease | momentum | confine |
| rhus | mitogenic | rohu | analogon | tuckahoe | decoupage |
| enough | collect | creature | hunt | columnist | merge |
| military | destroy | deficit | exclude | sensitivity | proposition |
| support | train | rapidly | confuse | planner | descent |
| news | climb | alternative | monster | streak | editorial |
| octroy | ouabain | cuticula | forepost | advertonal | pedium |
| organization | emotional | ultimate | outdoor | carrot | furious |
| close | conclusion | mechanism | transmission | jurisdiction | precedent |
| protect | muscle | incentive | convinced | forbid | astronot |
| prepare | governor | complicated | suspicion | fragment | spark |
| pikake | contrist | flute | greylag | arguendo | garefowl |
| discuss | except | pride | transaction | integrated | preacher |
| forget | issue | offense | equity | spy | governmental |
| maintain | dish | variation | strip | accurately | incidence |
| fish | waste | witness | chop | apology | circulate |
| fluerics | clowder | folleyer | punctum | barsom | hominal |
Test continued

| Level 7 | Level 8 | Level 9 | Level 10 | Level 11 | Level 12 |
|---------|---------|---------|----------|----------|----------|
| 7k-word band | 8k-word band | 9k-word band | 10k-word band | 11k-word band | 12k-word band |
| predominantly | elusive | succumb | persecution | pediatrician | twirl |
| complicate | extensively | executive | excavation | wreath | ancestry |
| murmur | coherent | meaningless | envy | divisive | gobble |
| controller | pave | flicker | deterioration | remedy | uncanny |
| hypostase | nosean | cloak | proeutectoid | madge | synusia |
| notorious | infectious | prelude | hamper | overlap | intrusive |
| compact | obesity | restructure | fluorescent | revert | cognition |
| saddle | pearl | scenery | cramped | stimulating | magistrate |
| considering | visibility | dwelling | cranberry | minced | refue |
| incus | cascabel | cohitre | ervil | orseille | merosome |
| chuckle | severity | synthesis | affinity | retarded | canister |
| situate | prolonged | aroma | aesthetics | precipitate | divergent |
| devastate | plight | mingle | warden | inspiring | espionage |
| skeleton | daunting | cloak | tenor | aquarium | concerted |
| manroot | civilite | demivol | sucupira | ailesterase | extrophy |
| retention | validate | accumulation | bestow | stoop | creepy |
| manifestation | capture | dissatisfaction | relinquish | beacon | tabloid |
| propel | impact | vengeance | persuasion | margarine | trickle |
| elevation | parliamentary | affective | monarchy | confound | deceptive |
| norleucine | pallilogy | dispersoid | trochalin | insilicate | biunal |
| unveil | discern | visualize | luminous | amnesty | intertwine |
| gradual | intentionally | glimpse | oppressive | expend | milieu |
| ammunition | rehearse | battalion | intimidation | wiggle | fidelity |
| pluck | resonate | mundane | dissipate | stash | squander |
| vervelle | endexine | rotenoid | trochantin | monorhinic | cessionaire |

Appendix B

**Lemma-Based Vocabulary Size Test (LVST)—Version B**

This test is designed to measure your written receptive vocabulary knowledge of English words.

| Level 1 | Level 2 | Level 3 | Level 4 | Level 5 | Level 6 |
|---------|---------|---------|---------|---------|---------|
| 1k-word band | 2k-word band | 3k-word band | 4k-word band | 5k-word band | 6k-word band |
| from | owner | gate | preference | revolutionary | congregation |
| like | positive | guilty | trick | trailer | surrender |
| still | alone | observation | register | sword | ward |
| put | hope | drag | appoint | bug | mandatory |
| porta | dungan | erythrism | twinspur | violescent | fonio |
| work | vote | territory | devote | thigh | charming |
| side | solution | negotiation | awful | manipulate | acute |
| follow | crowd | habit | colonial | cast | endangered |
| history | flight | fundamental | highlight | legislature | attain |
| curuba | cotylloid | pisote | petroleone | linaloe | etchant |
| offer | feed | commander | echo | patent | grind |
| free | cook | objective | entitle | stance | handling |

Please look at the words in the table. Some of these words are real English words and some are used as control words. Be aware that *Yes* responses to the control words are penalized. Please tick (✓) the words that you know or can use.

*Note: Your overall score will be affected severely by ticking control words.*

(continued)
| Level 1          | Level 2          | Level 3          | Level 4          | Level 5          | Level 6          |
|------------------|------------------|------------------|------------------|------------------|------------------|
| 1k-word band     | 2k-word band     | 3k-word band     | 4k-word band     | 5k-word band     | 6k-word band     |
| event            | colleague        | particular       | plead            | chunk            | parenting        |
| test             | expand           | clinical         | obstacle         | compliance       | prey             |
| halse            | sciara           | fibroma          | phyllade         | ruddle           | terna            |
| choose           | presidential     | valuable         | random           | toll              | utterly          |
| church           | shadow           | profile          | amendment        | constraint       | con              |
| lie              | notion           | adviser          | intelligent      | opt              | intimacy         |
| disease          | relief           | concentration    | residence        | rental            | span             |
| chicalote        | triazine         | eminente         | sampleite        | velutino          | scoldenore       |
| clearly          | typical          | founder          | counterpart       | favorable         | cube             |
| claim            | complain         | smell            | giant            | riot              | parallel         |
| respond          | beach            | comfort          | blend            | delight           | luxury           |
| chair            | faculty          | guide            | value            | sensor            | maximize         |
| slope            | roding           | dioptrics        | cynegetic        | histocyte         | banstickle       |

**Test continued**

| Level 7          | Level 8          | Level 9          | Level 10         | Level 11         | Level 12         |
|------------------|------------------|------------------|------------------|------------------|------------------|
| 7k-word band     | 8k-word band     | 9k-word band     | 10k-word band    | 11k-word band    | 12k-word band    |
| accountant       | skepticism       | breach           | longevity        | ensuing          | attributable     |
| hedge            | apparatus        | confide          | insignificant    | manure           | distorted        |
| coat             | intensely        | novelty          | freak            | pertinent        | dweller          |
| activate         | informant        | successive       | experimentation  | equitable        | gutter           |
| patelline        | epornitic        | propriation      | disseminule      | psychrophile     | cervelie         |
| heroic           | assault          | infected         | fend             | hegemony         | leukemia         |
| prone            | tribunal         | drastic          | enclave          | convergence      | articulation     |
| hammer           | crumb            | chill            | cartridge        | grope            | verification     |
| stripe           | focal            | converge         | expansive        | vain             | illumination     |
| cuticula         | gemauve          | converge         | foedatus         | pleonaste        | baldashin        |
| wage             | extraordinarily  | interception     | stray            | analogous        | saturate         |
| wrestle          | alternate        | spear            | replica          | accreditation    | restored         |
| deployment       | spike            | litter           | hastily          | dissatisfied     | bogus            |
| governance       | impatient        | cumulative       | orphanage        | quaint           | incorporation    |
| chellion         | yeatmanite       | molave           | frogbit          | incontrostrin    | ketoglutarate    |
| beneficiary      | preview          | flare            | ruthless         | disintegrate     | adversity        |
| spontaneous      | catastrophic     | thread           | tinkle           | entrust          | maternity        |
| diter            | collide          | absorption       | ashore           | gracefully       | offensive        |
| elicit           | lessen           | paranoid         | malpractice      | surplus          | coherence        |
| amgarn           | metasilicate     | scouse           | pectoralis       | motacilla        | aliesterase      |
| circular         | insult           | relish           | meager           | revitalize        | troupe           |
| paradise         | superiority     | groom            | disdain          | arch             | conscientious    |
| discriminate     | brighten         | freelance        | interpretive     | emigrate         | blush            |
| steroids         | painfully        | internally       | harness          | investing        | commotion        |
| buplever         | tylostyle        | slurb            | kantiara         | neral            | ascidiozoa       |
Declaration of Conflicting Interests
The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical Approval
All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Funding
The author received no financial support for the research, authorship, and/or publication of this article.

Informed Consent
Informed consent was obtained from all individual participants included in the study.

ORCID iD
Ahmed Masrai https://orcid.org/0000-0002-6778-5952

References
Aitchison, J. (2012). Words in the mind: An introduction to the mental lexicon. Wiley-Blackwell.
Alderson, J. C. (2005). Diagnosing foreign language proficiency: The interface between learning and assessment. Continuum.
Anderson, A., & Freebody, P. (1983). Reading comprehension and the assessment and acquisition of word knowledge. In B. Hutson (Ed.), Advances in reading/language research: A research annual (pp. 231–256). IAI Press.
Bauer, L., & Nation, P. (1993). Word families. International Journal of Lexicography, 6, 253–279.
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.
Brown, T. A. (2006). Confirmatory factor analysis for applied research. Guilford Press.
Cameron, L. (2002). Measuring vocabulary size in English as an additional language. Language Teaching Research, 6(2), 145–173.
Cohen, J. (1988). Statistical power analysis (2nd ed.). Erlbaum.
Davies, M. (2008). The Corpus of Contemporary American English: 320 million words, 1990–present. Retrieved from http://corpus.byu.edu/coca/
Eyckmans, J. (2004). Measuring receptive vocabulary size. LOT.
Field, A. (2009). Discovering statistics using SPSS (3rd ed.). Sage Publications Ltd.
Francis, W. N., & Kučera, H. (1982). Frequency analysis of English usage. Houghton Mifflin.
Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? Applied Linguistics, 11(4), 341–363.
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.
Harrington, M., & Carey, M. (2009). The on-line yes/No test as a placement tool. System, 37, 614–626.
Hermans, D. (2000). Word production in a foreign language. Unpublished dissertation, University of Nijmegen.
Hindmarsh, R. (1980). Cambridge English lexicon. Cambridge University Press.
Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes-no vocabulary test: Correction for guessing and response style. Language Testing, 19, 227–245.
Kempe, V., & MacWhirter, B. (1996). The cross-linguistic assessment of foreign language vocabulary learning. Applied Psycholinguistics, 17, 149–183.
Kremmel, B. (2016). Word families and frequency bands in vocabulary tests: Challenging conventions. TESOL Quarterly, 50(4), 976–987.
Kremmel, B., & Schmitt, N. (2016). Interpreting vocabulary test scores: What do various item formats tell us about learners’ ability to employ words? Language Assessment Quarterly, 13(4), 377–392.
Laufer, B. (1992). How much lexis is necessary for reading comprehension? IntH. Bejoint & P. Arnaud (Eds.), Vocabulary and applied linguistics (pp. 126–132). Macmillan.
Laufer, B., & Cobb, T. (2020). How much knowledge of derived words is needed for reading? Applied Linguistics, 41, 971–998. https://doi.org/10.1093/applin/amz051
Laufer, B., & Ravenhorst-Kalovski, G. (2010). Lexical threshold revisited: Lexical text coverage, learners’ vocabulary size and reading comprehension. Reading in a Foreign Language, 22(1), 15–30.
Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. Behavior Research Methods, 44, 325–343.
Masrai, A., & Milton, J. (2018). Measuring the contribution of academic and general vocabulary knowledge to learners’ academic achievement. Journal of English for Academic Purposes, 31, 44–57.
McLean, S. (2015). The vocabulary knowledge framework. Discussion Paper. Swansea University.
McLean, S., & Kramer, B. (2015). The creation of a new vocabulary placement tool. Shiken, 19(2), 1–11.
Meara, P. (1989). Matrix models of vocabulary acquisition. AILA Review, 6, 66–74.
Meara, P. (1992). EFL vocabulary tests. Centre for Applied Language Studies.
Meara, P. (1996). The vocabulary knowledge framework. Discussion Paper. Swansea University.
Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. Language Testing, 4(2), 142–154.
Meara, P., & Jones, G. (1990). Eurocentre’s vocabulary size test. Eurocentres.
Meara, P., & Milton, J. (2003). The Swansea levels test. Newbury Express.
Medical Outcomes Trust Scientific Advisory Committee. (1995). Instrument review criteria. Medical Outcomes Trust Bulletin, 3, 1–4.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (pp. 13–104). The Oryx Press.
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

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

Milton, J., & Treffers-Daller, J. (2013). Vocabulary size revisited: The link between vocabulary size and academic achievement. *Applied Linguistics Review, 4*(1), 151–172.

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, & M. Torreblanca-López (Eds.), *Further insights into non-native vocabulary teaching and learning* (pp. 83–98). Multilingual Matters.

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

Nation, I. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*, 59–82.

Nation, P. (1984). Vocabulary lists: Words, affixes and stems. English University of Wellington; English Language Institute.

Nation, P. (1990). *Teaching and learning vocabulary*. Heinle & Heinle Publishers.

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

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

Pallant, J. (2007). *SPSS survival manual*. Open University Press.

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

Pinchbeck, G. (2016). Developmental scales of L1 and L2 academic English vocabulary: Vocabulary test item difficulty indicates lexical sophistication and derivational morphology development. Paper presented at the American Association for Applied Linguistics conference, Orlando, FL.

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

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. Laufè (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 209–227). John Benjamins.

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

Roche, T., & Harrington, M. (2013). Recognition vocabulary knowledge as a predictor of academic performance in an English as a foreign language setting. *Language Testing in Asia, 3*(1), 1–13.

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

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

Schmitt, N., Nation, P., & Kremmel, B. (2019). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching, 53*(1), 109–120.

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.

Shillaw, J. (1999). *The Application of the Rasch model to Yes/No vocabulary tests* [Unpublished PhD thesis]. University of Wales Swansea.

Snellings, P., van Gelderen, A., & de Grolper, K. (2002). Lexical retrieval: An aspect of fluent second-language production that Can Be enhanced. *Language Learning, 52*, 723–754.

Thornndike, E. L., & Lorge, I. (1944). *The teacher’s word book of 30,000 words*. Teachers College, Columbia University.

Treffers-Daller, J., Parslow, P., & Williams, S. (2016). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics, 37*(4), 1–27.

van Zeelaland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics, 34*(4), 457–479.

Wang, Y., & Treffers-Daller, J. (2017). Explaining listening comprehension among L2 learners of English: The contribution of general language proficiency, vocabulary knowledge and metacognitive awareness. *System, 65*(Supplement C), 139–150.

Webb, S. (2013). Depth of vocabulary knowledge. In C. Chapelle (Ed.), *Encyclopaedia of Applied Linguistics* (pp. 1656–1663). Wiley-Blackwell.

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

Yule, G., Yanz, J. L., & Tsuda, A. (1985). Investigating aspects of the language learner’s confidence: An application of the theory of signal detection. *Language Learning, 35*, 473–488.

Zhang, X., Liu, J., & Ai, H. (2020). Pseudowords and guessing in second language reading. *Quarterly, 63*(1), 5–31.