OPEN ACCESS

Vocabulary Learning and Instruction
ISSN 2981-9954
Volume 13, Number 2 (2024)
https://doi.org/10.29140/vli.v13n2.1604

Castledown

Key Issues and Considerations in Measuring Vocabulary Growth: A Methodological Overview

Abdullah Albalawi
Saudi Electronic University, King Fahad Road, Tabuk, Saudi Arabia
a.balawi@seu.edu.sa

Abstract

Despite the substantial expansion in vocabulary research since the 1980s, we still know very little about how vocabulary develops over time and what factors influence this development. This methodological overview discusses key issues and considerations in vocabulary breadth growth assessment to help advance research in this area. The report begins by discussing general issues in vocabulary assessment such as sampling rate and the effect of cognates. This is followed by an overview and an evaluation of common vocabulary breadth tests. The report ends with recommendations for choosing vocabulary tests for vocabulary growth research.

Keywords: vocabulary, growth, development, method, assessment.

Introduction

Despite the substantial expansion in vocabulary research since the 1980s (Laufer, 2009; Meara, 2002), we still know very little about how vocabulary develops over time and what factors influence this development (Pellicer-Sánchez, 2019; Webb & Nation, 2017). Webb and Nation (2017, p. 68) state that "Surprisingly, there are relatively few studies of L2 vocabulary growth; and questions such as 'How many words should

Data Availability Statement: All relevant data are within this paper.

be learned per week/per year/during a course?' remain unanswered". Pellicer-Sanchez (2019) echoed this call, emphasizing the need for more longitudinal research on vocabulary growth.

Tracking vocabulary knowledge development over time can provide key insights to research on vocabulary knowledge development. A commonly stated observation within vocabulary research is the fact that the field has yet to develop an overall theory of vocabulary knowledge development (Schmitt, 2019). Research that tracks how vocabulary develops over time can provide empirical data against which theoretical assumptions can be tested. Additionally, research on vocabulary growth can have practical implications. For example, it can provide benchmarks for language programs to evaluate the vocabulary growth rate of their learners and whether they are progressing as expected or whether an intervention is needed. Moreover, it can help textbook developers make informed decisions regarding the number of target vocabulary to include in the textbooks (Milton, 2009). Milton and Hopwood (2022) note that many EFL textbooks lack a principled approach to the type and amount of vocabulary to include. For example, studies report that a proportion of important vocabulary (high frequency words) is missing from EFL textbooks with estimates ranging from 30% (in the Success coursebook series; Eldridge & Neufield, 2009) to 15% (Saudi EFL textbooks; Alsaif & Milton, 2012). Despite its importance, this line of inquiry remains relatively under-researched (Dóczi & Kormos, 2015; Pellicer-Sánchez, 2019; Webb & Nation, 2017). This methodological overview discusses key issues and considerations in vocabulary growth assessment to help advance research in this area. The report begins with an overview of general vocabulary assessment issues. It then moves on to discuss common vocabulary breadth tests. Finally, the report concludes with recommendations for selecting vocabulary tests for research on vocabulary breadth growth.

Measuring Vocabulary Knowledge

Vocabulary testing is important for vocabulary research, which is evident in the fact that vocabulary tests are among the most used research instruments in vocabulary research (Durrant et al., 2022). Whole books (Milton, 2009; Read, 2000), book sections (Durrant et al., 2022; Nation, 2022; Schmitt, 2020; Webb, 2019) and several articles (e.g., Read & Chapelle, 2001; Schmitt et al., 2019; Stoeckel et al., 2020) have been written on vocabulary testing to help move the field forward.

Given the difficulty in measuring all aspects of vocabulary knowledge at once (e.g., measuring knowledge of meaning, collocation and derivation in one test; Read, 2000), test developers tend to focus on one or a few aspects of vocabulary knowledge when designing tests (Laufer & Goldstein, 2004; Webb et al., 2017). Vocabulary tests can be classified according to three types of distinctions. The first is the distinction between tests that measure vocabulary *breadth* and tests that measure vocabulary *depth*. Tests that focus on breadth give estimates of how many words learners know by measuring the form-meaning component (e.g., form is provided, and learners supply the meaning). Tests focusing on depth tell us how well these words are known by measuring the other vocabulary knowledge components (e.g., asking learners not only to provide a word meaning, but also other components such as its collocations or associations). The second distinction is between *receptive* and *productive* vocabulary knowledge tests.

Table 1 *Vocabulary Knowledge of Form-Meaning Test Types*

	Meaning Recognition	Form Recognition	Meaning Recall	Form Recall
Provided	Form	Meaning	Form	Meaning
Tested	Meaning recognition	Form recognition	Meaning recall	Form recall
Example	1-Car a-furniture b-vehicle c-container	1-A type of vehicle a-car b-chair c-spoon	1-license …………..	1-a permit to use or own something 1…….

Receptive knowledge tests assess learners' ability to recognize the meaning of a word in reading or listening while productive knowledge tests test learners' ability to use a word in speaking or writing (Nation, 2022; Schmitt, 2010). The final distinction is between tests that measure *recognition* and *recall* knowledge of words. Recognition tests require learners to choose the correct form or meaning of a word, while recall tests require learners to retrieve from memory a word meaning or form. Focusing on written vocabulary breadth, the combination of receptive/productive and recognition/recall in Table 1 provides an overview of the four possibilities of vocabulary knowledge tests (Laufer & Goldstein, 2004; Schmitt, 2020).

Before exploring the vocabulary knowledge tests that can be used in longitudinal vocabulary research, it is important to review some key considerations in vocabulary assessment. This is important because some vocabulary tests have certain design issues and some have not been properly validated before publication (Durrant et al., 2022; Schmitt et al., 2019). The following sections aim to provide an evaluation of vocabulary tests and discuss their strengths and weaknesses.

Key Concepts and Issues in Vocabulary Assessment

Vocabulary tests, like any other language test, need to meet three key criteria before they can be used: validity, reliability, and practicality (Bachman, 1990; Bachman & Palmer, 1996). Validity in general refers to the key condition that a test needs to measure what it is supposed to measure and minimize influence from irrelevant factors. For example, if the intended construct of measurement is productive vocabulary knowledge, then receptive knowledge should not interfere significantly with the measurement. Reliability refers to consistency and stability in measurement. In other words, a test should give similar results if for example it was taken multiple times in the same session by the same learner. Similarly, if a test has two versions, they need to give similar results if they were taken by the same learner on the same day. Finally, practicality refers to the condition of efficiency in that a test for example should not take too much resource to administer and score (Nation, 2022). This is why, for example, most vocabulary tests focus on one aspect of vocabulary knowledge because attempting to test all aspects reliably would probably take too much time to administer. The three criteria of validity, reliability and practicality should be taken into consideration when evaluating the different vocabulary tests available. In addition to the broader language testing considerations, there are more vocabulary-focused issues that need to be considered, including the number of vocabulary aspects to test, item format, sampling rate and the influence of cognates (Durrant et al., 2022; Stoeckel et al., 2020).

Recognition or Recall?

Measuring knowledge of vocabulary breadth using recognition tests has been criticized on the grounds that they tend to overestimate the number of words learned due to random guessing (Gyllstad et al., 2015; Stoeckel et al., 2020). However, the opposite might happen if vocabulary knowledge is measured using only recall tests since they tend to underestimate the number of words learned (Kremmel & Schmitt, 2016). A more serious issue with recognition tests relates to ecological validity in that they might not reflect receptive vocabulary knowledge reliably (Gyllstad et al., 2015; Kremmel & Schmitt, 2016; Stewart, 2014). When learners use language receptively (reading or listening), they are not offered a list of meaning choices to choose from, but they must recall word meaning from the mental lexicon. How representative recognition tests are of receptive vocabulary knowledge remains open for further research (Stewart et al., 2021; Stoeckel et al., 2020; Webb, 2021).

Sampling Rate

Given the impracticality of testing all higher frequency words in a vocabulary levels test or all the words known by learners in vocabulary size tests (which both tend to be in the thousands), test developers normally resort to sampling 10 to 40 words from each frequency band and test these words only (Gyllstad et al., 2015; Laufer & Goldstein, 2004; Webb et al., 2017). When learners answer most of these words correctly, it is assumed that they know the majority of the other words in the frequency band. These assumptions are based on the finding that learners tend to learn more frequent words (e.g., *house*) before learning less frequent words (e.g., *dwelling*). Based on this, it has been hypothesized that if learners know a word in one frequency band (e.g., *expensive* from the first 1000 band) there is a good chance that they know the other words from the same frequency band (e.g., *good*, *happy*, *hot*). A key issue here is sampling rate or how many words should be tested from a frequency band to be deemed representative of mastery of the majority of words in that frequency band (Gyllstad et al., 2015; Stoeckel et al., 2020). Vocabulary tests vary between as little as 5 words per frequency band to as high as 40. One possible recommendation is 'the more the better', however, practicality would soon become an issue (Durrant et al., 2022). A more practical and seemingly sufficient threshold is 30 words per frequency band (Gyllstad et al., 2015, 2021). In Gyllstad et al. (2021) 103 Japanese EFL learners were tested on all the words in a frequency band (3000 band) using meaning recall and recognition tests. Using bootstrapping[1], they compared tests with 5, 10, 20, 30,

[1] Bootstrapping is "a type of robust statistic that simulates how a study would be replicated by resampling from a population." (LaFlair et al., 2015, p. 46).

40, 50, 60, 100 and 200 items to the students' actual test scores. They found that the mismatch between the bootstrap samples and the actual test scores declines as test items increase. The percentage of difference was highest with a sampling rate of five items (50% for recall item tests and 20% for recognition item tests) and least with the 200-item test (10% for recall test and 5% for recognition). More importantly, they found that the curve starts to flatten out after the 30-item threshold. Based on this, they recommend a sampling rate of 30 words per frequency band for both recognition and recall vocabulary tests.

Monolingual or Bilingual Tests?

When language learners take vocabulary tests, they bring with them their L1 resources which can influence test scores. Two main areas have been investigated in this regard: the role of translation and cognates (Durrant et al., 2022; Read, 2019). Translating vocabulary tests to learners' L1 (i.e., creating bilingual vocabulary tests) has been supported by Nation (2022) since the 1990s on the grounds that this might minimize the influence from factors other than vocabulary knowledge (e.g., knowledge of relative clauses, see: Nguyen & Nation, 2011) which should enhance the construct validity of the test. Following Nation's recommendation, a number of bilingual vocabulary tests have been developed for several languages such as Vietnamese (Nguyen & Nation, 2011) and Persian (Karami, 2012). Elgort (2013) provided evidence for Nation's recommendation when she compared a monolingual vocabulary test with a Russian bilingual vocabulary test. 121 intermediate EFL learners took both tests (70 items in each) and their results showed significantly higher scores (32.97) on the bilingual test than the monolingual one (29.61). Her findings suggest that giving a monolingual test can significantly underestimate the vocabulary knowledge of learners by up to 672 word-families. Thus, bilingual vocabulary tests might be more reliable measures of vocabulary knowledge than monolingual tests. The use of bilingual tests however can be challenging when learners in a classroom have different L1 backgrounds.

The Effect of Cognates and Loanwords

The second area where the role of L1 was examined is the effect of *cognates* on vocabulary test scores. Cognates or loanwords are words that share a similar sound and meaning in two languages (Laufer & McLean, 2016). For example, the Spanish word *persona* and the English word *person* are considered cognates because they have a similar phonological form and meaning across the two languages. The two terms cognates and loanwords are often used interchangeably. However, when talking about two genetically unrelated languages such as Arabic and English, the term loanwords might be more appropriate since these languages do not share a common ancestor (Laufer & McLean, 2016). One of the most common areas where other languages have borrowed words from English is in the area of technology. For example, the Arabic words *televizion* (television), *fedio* (video) and *combuter* (computer) are loanwords that were borrowed from English.

When it comes to vocabulary testing, cognates and loanwords pose a challenge for test developers and researchers. Cognates and loanwords tend to be answered more correctly than non-cognates and non-loanwords (Allen, 2018, 2019a, 2020; Canning et al., 2024; De Wilde, 2023; Elgort, 2013; Laufer & McLean, 2016). In itself this is not an issue given that cognates and loanwords are part of the learners' lexicon and they should be represented in the vocabulary knowledge estimates (Nation & Webb, 2011). However, it might become a problem when the proportion of cognates and loanwords in a test is not representative of their proportion in the language (Cobb, 2000; Laufer & Levitzky-Aviad, 2018). This can lead to either overestimation or underestimation of vocabulary knowledge. For example, Elgort (2013) found that the proportion of English-Russian cognates in a vocabulary test was 34% which is higher than the 27% proportion found in the wordlist which the test items were sampled from. This can lead to overestimation in vocabulary knowledge (Allen, 2019a; Elgort, 2013). One solution is to develop a customized vocabulary test for a homogenous group of EFL learners who share a common L1 which takes into account the accurate proportion of cognates (Peters et al., 2019). The situation is more complicated when a group of learners have different L1s (Laufer & McLean, 2016), and no solution appears to be viable that ensures an accurate representation of cognates in the vocabulary knowledge estimates in this case.

For languages that are not genetically related to English and do not have many English loanwords such as Hebrew, the effect of these words seems nonsignificant. Laufer and Levitzky-Aviad (2018) examined how the presence of English-Hebrew loanwords affected the vocabulary test scores of 303 Hebrew EFL learners with three levels of proficiency. The learners took tests with varying numbers of loanwords, including tests with no loanwords, tests with a representative number of loanwords and tests with a random number of loanwords. These tests covered four aspects of written vocabulary knowledge: word form recall, word meaning recall, word form recognition, and word meaning recognition. The results showed that the impact of loanwords on test scores varied depending on the specific format of the test and the proficiency levels of the learners. The key finding is that the score increase from the version of the test with representative loanwords to the version with random loanwords was minimal, and the differences in the effect size were very small. Therefore, overall, Laufer and Levitzky-Aviad (2018) findings suggest suggests that loanwords in vocabulary tests may not significantly affect the accuracy of measuring true vocabulary knowledge. Overall, although cognates and loanwords have a significant facilitating effect that tends to inflate English vocabulary test scores, the magnitude of the effect seems to depend on the L1 of the learners. The influence appears to be minimal for some languages such as Hebrew (Laufer & Levitzky-Aviad, 2018) and larger for genetically related languages such as French (Cobb, 2000) and languages with more borrowings from English such as Japanese (Allen, 2019a, 2019b; Daulton, 2008).

In summary, like any language test, vocabulary tests need to meet the criteria of validity, reliability, and practicality. In addition to these criteria, vocabulary researchers need to be aware of other factors that might have an influence on vocabulary testing such as item format, sampling rate, translation and cognates. Having reviewed these key concepts and issues, the vocabulary tests discussed in the next section can be better evaluated and critically examined.

What Vocabulary Tests can be Used in Vocabulary Breadth Growth Research?

Since the 1980s, several vocabulary breadth tests have been developed. The following list shows some of the commonly used tests of vocabulary form-meaning knowledge:

- Vocabulary Levels Test (Nation, 1983, 1990; Schmitt et al., 2001; Webb et al., 2017)
- Checklist tests (Meara, 1992; Meara & Jones, 1988)
- Computer Adaptive Test of Size & Strength (Aviad-Levitzky et al., 2019; Laufer & Goldstein, 2004)
- Vocabulary Size Test (Nation & Beglar, 2007)

Earlier tests (Laufer & Goldstein, 2004; Nation, 1983, 1990; Schmitt et al., 2001) relied on wordlists that were based on small and potentially outdated corpora such as the General Service List (West, 1953) to determine word frequency (e.g., the list includes words such as scold and coward which are not likely to be regarded as high frequency words today; Webb & Sasao, 2013). With the advent of computerized and large corpora such as the British National Corpus and the Corpus of Contemporary American English, more accurate and up-to-date wordlists were created (Nation, 2006) which later tests (Aviad-Levitzky et al., 2019; Nation & Beglar, 2007; Webb et al., 2017) relied on.

The Vocabulary Levels Test (VLT; Nation, 1983, 1990; Schmitt et al., 2001) is possibly the most widely used tests of learners' vocabulary knowledge (Read, 2000; Schmitt, 2020). The early versions measure learners' receptive knowledge of words at four frequency levels (2000, 3000, 5000 and 10,000) and their knowledge of academic words. The Updated Vocabulary Levels Test (UVLT; Webb et al., 2017) differs from previous VLTs in that it uses updated wordlists and measures every 1000 frequency level from the first 5000 words (previous VLTs skipped the first and fourth frequency levels). This comes at the expense of excluding the 10,000-frequency level and the academic vocabulary part. One of the test's strengths lies in that it has a higher sampling rate (30 items per 1000 frequency level) compared to other tests, which can provide more accurate vocabulary knowledge estimates (Gyllstad et al., 2015, 2021). Two versions of the UVLT were initially validated by Webb et al. (2017) on 250 university students from three countries (China, Japan and Spain). The results suggest that the test is a valid (e.g., the test difficulty increases as words' frequency decreases) and reliable measure of written receptive knowledge (Rasch item and person reliability = .96).

Checklist tests (Meara, 1992; Meara & Jones, 1988) measure learners' vocabulary size by presenting sample words from different frequency levels and asking learners to indicate the words they think they know. These tests differ from other tests in that learners are not required to demonstrate their knowledge of form-meaning link, which can be problematic since some learners might overestimate their lexical knowledge. One common convention to overcome this shortcoming is by including nonwords (i.e., made-up words) to adjust the overall score for possible overestimation. Some checklist tests, such as X-Lex (Meara & Fitzpatrick, 2000), have no validation records.

Most vocabulary tests have the limitation of testing only one aspect of vocabulary knowledge (e.g., meaning recognition only). However, The Computer Adaptive Test of Size & Strength (CATSS; Aviad-Levitzky et al., 2019; Laufer & Goldstein, 2004) overcomes this shortcoming by testing all four aspects of the form-meaning components of vocabulary knowledge. It tests the strength of the form-meaning connection on the basis of four formats: meaning recognition, form recognition, meaning recall and form recall. The new version of the CATSS uses more updated word lists and measures learners' total vocabulary knowledge by sampling 140 words from the first 14 1000 frequency bands. It was validated on 453 university students and appears to be valid and reliable test in this context (overall test Cronbach's Alpha = .98). The test's advantage of testing words across four formats comes at a cost since it suffers from a low sampling rate of only 10 items per 1000 band, which some consider insufficient to represent the whole frequency level (Gyllstad et al., 2015; Schmitt et al., 2001). Moreover, the low sampling rate means that this test may not be appropriate for testing vocabulary growth longitudinally, since newly learned words may be missed in the items chosen for testing.

The Vocabulary Size Test (VST; Nation & Beglar, 2007) measures learners' total vocabulary knowledge by sampling 10 words from each of the 1st 1000 frequency bands up until the 14th with a total of 140 items. It is a multiple-choice test that measures in particular the written receptive knowledge of the form-meaning link; thus, it does not provide information about the other vocabulary knowledge components. The VST has been suggested to be a reliable and accurate measure of learners' receptive knowledge of the most frequent 14,000 words (Beglar, 2010). However, the same low sampling issue and its implications in the CATTS might also apply to the VST.

This section provided an overview of common vocabulary breath tests. The list of tests discussed here is not exhaustive, and there are other vocabulary breadth tests that have been developed that were not discussed due to space, such as the New Vocabulary Levels Test (NVLT; McLean & Kramer, 2015) and the New General Service List Test (Stoeckel & Bennett, 2015). The following section provides some suggestions that might help in choosing a test to measure vocabulary breadth growth.

Key Considerations When Measuring Vocabulary Growth

A key question here is how what has been discussed so far relates to the measuring of vocabulary growth. When tracking vocabulary growth, vocabulary size tests[2] such as the VST and CATTS have the advantage of covering a wide range of frequency bands (1000–14,000) which can increase the likelihood of detecting growth. Increasing the frequency bands covered by a test usually comes at the expense of reducing the sampling rate in these tests (e.g., 10 words per 1000 frequency band in the VST and CATTS) which can decrease the likelihood of detecting vocabulary growth. Recent vocabulary levels tests (UVLT and NVLT) have the advantage of a higher sampling rate (UVLT = 30 per 1000 frequency band, NVLT = 24) but have the limitation of less coverage (e.g., both the UVLT and NVLT do not cover words beyond the most

[2] Excluding vocabulary size tests where learners are not required to demonstrate their knowledge such checklist tests.

frequent 5000 words). To give an example, suppose a language learner in a vocabulary growth study has learned several words in the 7000 frequency band, these words will not be detected in the vocabulary levels tests since this band is not covered in the tests but they might be detected in vocabulary size tests which cover this band. In terms of the sampling rate, if a learner learned new words for instance in the 3000 frequency band, these words would be more likely detected in vocabulary levels tests since they tend to have a higher sampling rate than vocabulary size tests.

This raises the question of which types of tests (levels or size) are more appropriate for research on vocabulary growth. One key factor that might help in making this decision is learners' proficiency. For more advanced learners, vocabulary size tests are more appropriate since learners are more likely to learn mid-frequency (3000–9000) and perhaps low frequency vocabulary (beyond 9000 words). For beginners with a small vocabulary size, sampling rate might be more important than coverage since they are less likely to learn words beyond the most frequent 5000 words. In fact, testing beginners on lower frequency bands might increase random guessing and hence overestimate vocabulary growth (Beglar, 2010; Elgort, 2013; Mclean et al., 2015; Stewart, 2014). Therefore, vocabulary levels tests such as the UVLT might be more appropriate for beginners than vocabulary size tests when tracking vocabulary growth.

The term *beginners* has been used somewhat vaguely here. This is because it is difficult to specify with confidence exactly when vocabulary levels tests are no longer appropriate to be used as vocabulary growth tests. One rule of thumb is to allow for two frequency bands beyond learners' current level (a rule originally suggested to minimize random guessing; Beglar, 2010; Elgort, 2013). Based on this, vocabulary levels tests might be sufficient for learners with a vocabulary size of up to 3000 word-families. It should be noted that as learners' vocabulary size increases, the likelihood of missing newly learned words starts to increase in vocabulary levels tests. Therefore, the UVLT is likely to be sufficient for learners with a vocabulary size of 1000 word-families or lower, but as their vocabulary size increases, the confidence in the growth estimates starts to decrease as they are more likely to learn words beyond the 5000-word-families limit.

When testing the vocabulary growth of a group of homogenous learners who share the same L1, then a bilingual vocabulary test - if any exists- might give more accurate estimates as discussed earlier (Elgort, 2013). Similarly, testing learners with the same L1 enables more control over the effect of cognates. If learners' L1 shares many cognates or loanwords with English (e.g., French and Japanese), some sort of adjustments might be needed to make sure growth estimates will not be over or under-estimated (Allen, 2019a; Elgort, 2013). This can involve using a test with a representative proportion of cognates and loanwords if one exists (Peters et al., 2019) or modifying an existing test (Jordan, 2012). For learners whose L1 does not share many cognates or loanwords with English such as Hebrew, the distortion coming from these words seems negligible (Laufer & Levitzky-Aviad, 2018).

There is little research on the influence of test format (recognition and recall) on vocabulary growth (Dóczi & Kormos, 2015). However, previous cross-sectional research shows that learning vocabulary to the recall level is more difficult than learning to the recognition level (González-Fernández & Schmitt, 2019; Laufer & Goldstein, 2004).

What this entails to vocabulary growth is that recognition item tests might be more sensitive to initial development in vocabulary and could lead to higher growth estimates compared to recall tests. As discussed before, which is a more accurate predictor of vocabulary knowledge is open for further investigation. For now, using both if possible should give more insights into longitudinal vocabulary development.

Conclusion

Schmitt (2010, p. 156) points out that "vocabulary learning is longitudinal and incremental in nature, and only research designs with a longitudinal element can truly describe it". As discussed in the introduction of this report, vocabulary growth research is important and has key implications for vocabulary theory and practice. This methodological review aimed to advance research in this area by highlighting key issues and considerations in vocabulary breadth assessment for longitudinal research. General issues, such as test format, sampling rate, and cognates, were discussed, along with an evaluation of common vocabulary breadth tests. These recommendations should contribute to improving the accuracy and reliability of vocabulary growth assessment.

References

Allen, D. (2018). Cognate frequency and assessment of second language lexical knowledge. *International Journal of Bilingualism*, *23*(5), 1121–1136. https://doi.org/10.1177/1367006918781063

Allen, D. (2019a). Cognate frequency predicts accuracy in tests of lexical knowledge. *Language Assessment Quarterly*, *16*(3), 312–327. https://doi.org/10.1080/15434303.2019.1635134

Allen, D. (2019b). The prevalence and frequency of Japanese-English cognates: Recommendations for future research in applied linguistics. *International Review of Applied Linguistics in Language Teaching*, *57*(3), 355–376. https://doi.org/10.1515/iral-2017-0028

Allen, D. (2020). An overview and synthesis of research on English loanwords in Japanese. *Vocabulary Learning and Instruction*, *9*(1), 33–50. https://doi.org/10.1177/1367006918781063

Alsaif, A., & Milton, J. (2012). Vocabulary input from school textbooks as a potential contributor to the small vocabulary uptake gained by English as a foreign language learners in Saudi Arabia. *The Language Learning Journal*, *40*(1), 21–33. https://doi.org/10.1080/09571736.2012.658221

Aviad-Levitzky, T., Laufer, B., & Goldstein, Z. (2019). The new computer adaptive test of size and strength (CATSS): Development and validation. *Language Assessment Quarterly*, *16*(3), 345–368. https://doi.org/10.1080/15434303.2019.1649409

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

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

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing*, *27*(1), 101–118. https://doi.org/10.1177/0265532209340194

Canning, D., McLean, S., & Vitta, J. P. (2024). Relative complexity in a model of word difficulty: The role of loanwords in vocabulary size tests. *Studies in Second Language Learning and Teaching*. Advanced Online Publication. https://doi.org/10.14746/ssllt.38492

Cobb, T. (2000). One size fits all? Francophone learners and English vocabulary tests. *The Canadian Modern Language Review*, *57*(2), 295–324. https://doi.org/10.3138/cmlr.57.2.295

Daulton, F. E. (2008). *Japan's built-in lexicon of English-based loanwords*. Multilingual Matters.

De Wilde, V. (2023). The auditory picture vocabulary test for English L2: A spoken receptive meaning-recognition test intended for Dutch-speaking L2 learners of English. *Language Teaching Research*, 1–31. https://doi.org/10.1177/13621688221147462

Dóczi, B., & Kormos, J. (2015). *Longitudinal developments in vocabulary knowledge and lexical organization*. Oxford University Press.

Durrant, P. L., Siyanova-Chanturia, A., Kremmel, B., & Sonbul, S. (2022). *Research methods in vocabulary studies*. John Benjamins Publishing Company. https://doi.org/10.1075/rmal.2

Eldridge, J., & Neufield, S. (2009). The graded reader is dead, long live the electronic reader. *Reading*, *9*(2), 224–244. http://www.readingmatrix.com/articles/sept_2009/eldridge_neufeld.pdf

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing*, *30*(2), 253–272. https://doi.org/10.1177/0265532212459028

González-Fernández, B., & Schmitt, N. (2019). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*, *41*(4), 481–505. https://doi.org/10.1093/applin/amy057

Gyllstad, H., McLean, S., & Stewart, J. (2021). Using confidence intervals to determine adequate item sample sizes for vocabulary tests: An essential but overlooked practice. *Language Testing*, *38*(4), 558–579. https://doi.org/10.1177/0265532220979562

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. https://doi.org/10.1075/itl.166.2.04gyl

Jordan, E. (2012). Cognates in vocabulary size testing - a distorting influence? *Language Testing in Asia*, *2*(3), 5. https://doi.org/10.1186/2229-0443-2-3-5

Karami, H. (2012). The development and validation of a bilingual version of the vocabulary size test. *RELC Journal*, *43*(1), 53–67. https://doi.org/10.1177/0033688212439359

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. https://doi.org/10.1080/15434303.2016.1237516

LaFlair, G. T., Egbert, J., & Plonsky, L. (2015). A practical guide to bootstrapping descriptive statistics, correlations, t tests, and ANOVAs. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 46–77). Routledge.

Laufer, B. (2009). Second language vocabulary acquisition from language input and from form-focused activities. *Language Teaching*, *42*(2), 341–354. https://doi.org/10.1017/S0261444809005771

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., & Levitzky-Aviad, T. (2018). Loanword proportion in vocabulary size tests. *ITL - International Journal of Applied Linguistics*, *169*(1), 95–114. https://doi.org/10.1075/itl.00008.lau

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

McLean, S., & Kramer, B. (2015). The creation of a new vocabulary levels test. *Shiken*, *19*(2), 1–11. http://teval.jalt.org/sites/teval.jalt.org/files/Shiken_19-02_Complete-1.pdf#page=26

Mclean, S., Kramer, B., & Stewart, J. (2015). An empirical examination of the effect of guessing on vocabulary size test scores. *Vocabulary Learning and Instruction*, *4*(1), 1–10. https://doi.org/10.7820/vli.v04.1.mclean.et.al

Meara, P. (1992). *EFL vocabulary tests*. ERIC Clearinghouse.

Meara, P. (2002). The rediscovery of vocabulary. *Second Language Research*, *18*(4), 393–407. https://doi.org/10.1191/0267658302sr211xx

Meara, P., & Fitzpatrick, T. (2000). Lex30: An improved method of assessing productive vocabulary in an L2. *System*, *28*(1), 19–30. https://doi.org/10.1016/S0346-251X(99)00058-5

Meara, P., & Jones, G. (1988). Vocabulary size as a placement indicator. In P. Grunwell (Ed.), *Applied Linguistics in Society.* (pp. 80–87). CIL.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Multilingual Matters. https://doi.org/10.21832/9781847692092

Milton, J., & Hopwood, O. (2022). *Vocabulary in the foreign language curriculum*. Routledge. https://doi.org/10.4324/9781003278771

Nation, P. (1983). Testing and teaching vocabulary. *Guidelines*, *5*(1), 12–25.

Nation, P. (1990). *Teaching and Learning Vocabulary*. Heinle and Heinle.

Nation, P. (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, P. (2022). *Learning vocabulary in another language*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524759

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*. https://doi.org/10.1177/0265532209340194

Nation, P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Cengage Learning.

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

Pellicer-Sánchez, A. (2019). Examining second language vocabulary growth: Replications of Schmitt (1998) and Webb & Chang (2012). *Language Teaching*, *52*(4), 512–523. https://doi.org/10.1017/S026144481800037X

Peters, E., Velghe, T., & Van Rompaey, T. (2019). The VocabLab tests. *ITL - International Journal of Applied Linguistics*, *170*(1), 53–78. https://doi.org/10.1075/itl.17029.pet

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732942

Read, J. (2019). Key issues in measuring vocabulary knowledge. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 545–560). Routledge.

Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing*, *18*(1), 1–32. https://doi.org/10.1177/026553220101800101

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

Schmitt, N. (2019). Understanding vocabulary acquisition, instruction, and assessment: A research agenda. *Language Teaching*, *52*(2), 261–274. https://doi.org/10.1017/S0261444819000053

Schmitt, N. (2020). *Vocabulary in language teaching*. Cambridge University Press.

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*, 109–120. https://doi.org/10.1017/S0261444819000326

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.

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

Stewart, J., Stoeckel, T., McLean, S., Nation, P., & Pinchbeck, G. G. (2021). What the research shows about written receptive vocabulary testing: A reply to Webb. *Studies in Second Language Acquisition*, *43*(2), 462–471. https://doi.org/10.1017/S0272263121000437

Stoeckel, T., & Bennett, P. (2015). A test of the new general service list. *Vocabulary Learning and Instruction*, *4*(1), 1–8.

Stoeckel, T., McLean, S., & Nation, P. (2020). Limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition*, 1–23. https://doi.org/10.1017/S027226312000025X

Webb, S. (2019). The Routledge handbook of vocabulary studies. In *The Routledge Handbook of Vocabulary Studies*. Routledge. https://doi.org/10.4324/9780429291586

Webb, S. (2021). A different perspective on the limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition*, *43*(2), 454–461. https://doi.org/10.1017/S0272263121000449

Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.

Webb, S., & Sasao, Y. (2013). New directions in vocabulary testing. *RELC Journal*, *44*(3), 263–277. https://doi.org/10.1177/0033688213500582

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test. *ITL International Journal of Applied Linguistics*, *168*(1), 33–69. https://doi.org/10.1075/itl.168.1.02web

West, M. (1953). *A general service list of English words*. Longman, Green & Co.