Studies in Second Language Acquisition (2020), 1–23
doi:10.1017/S027226312000025X

State-of-the-Scholarship Article

LIMITATIONS OF SIZE AND LEVELS TESTS OF WRITTEN RECEPTIVE VOCABULARY KNOWLEDGE

Tim Stoeckel *

University of Niigata Prefecture

Stuart McLean

Momoyama Gakuin University

Paul Nation

Victoria University of Wellington

Abstract
Two commonly used test types to assess vocabulary knowledge for the purpose of reading are size and levels tests. This article first reviews several frequently stated purposes of such tests (e.g., materials selection, tracking vocabulary growth) and provides a reasoned argument for the precision needed to serve such purposes. Then three sources of inaccuracy in existing tests are examined: the overestimation of lexical knowledge from guessing or use of test strategies under meaning-recognition item formats; the overestimation of vocabulary knowledge when receptive understanding of all word family members is assumed from a correct response to an item assessing knowledge of just one family member; and the limited precision that a small, random sample of target words has in representing the population of words from which it is drawn. The article concludes that existing tests lack the accuracy needed for many specified testing purposes and discusses possible improvements going forward.

Two commonly used test types to assess vocabulary knowledge for the purpose of reading are size and levels tests. Despite a flurry of research activity on the development and validation of such instruments in recent years (e.g., Beglar, 2010; Gyllstad et al., 2015; Laufer & Goldstein, 2004; Stewart & White, 2011; Stoeckel et al., 2016; Webb et al.,

We would like to extend special thanks to Phil Bennett, Dale Brown, and Brandon Kramer for their kind and useful suggestions for improving this article.

* Correspondence concerning this article should be addressed to Tim Stoeckel, University of Niigata Prefecture, 471 Ebigase, Higashi-ku, Niigata City, Niigata 950-8680, Japan. E-mail: stoeckel@unii.ac.jp

2017), there are limitations as to how accurately existing tests measure lexical knowledge, and this limits the ways in which scores ought to be interpreted. This article examines three sources of inaccuracy in tests of written receptive vocabulary knowledge. First is the overestimation of lexical knowledge that can result from guessing or use of test strategies under multiple-choice and matching item formats (Kremmel & Schmitt, 2016; Stewart & White, 2011; Stoeckel et al., 2019; Zhang, 2013). Second is the overestimation of vocabulary knowledge that can occur when receptive understanding of all members of a word family is assumed from a correct response to an item assessing knowledge of just one family member (McLean, 2018; Mochizuki & Aizawa, 2000; Sasao & Webb, 2017; Stoeckel et al., 2018a; Ward & Chuenjundaeng, 2009). Third is the limited precision that a small, random sample of target words has in representing the population of words from which it is drawn (Gyllstad et al., 2019).

In our discussion of these three issues, we will focus exclusively on tests that assess understanding of the form-meaning link, or the pairing in the mental lexicon of a word form with one of its meaning senses. We do not consider yes/no tests (e.g., Meara & Buxton, 1987). Such instruments assess the ability to recognize that a word form belongs to the lexicon of a language, which is an important step in the development of lexical knowledge, but form-recognition alone is not enough to make sense of a language. Moreover, reading proficiency is more closely related to scores on vocabulary tests that assess form-meaning linkage than scores on yes/no tests (McLean et al., 2020).

PURPOSES OF SIZE AND LEVELS TESTS

To evaluate whether lost accuracy influences the effectiveness of size or levels tests, we must have a clear idea of the purposes for which they were made (Schmitt et al., 2019). Size tests are generally designed to estimate the total number of words learners know, and levels tests to determine whether learners have mastery of specified word bands. Test makers have also stated that such size estimates or mastery statements could be used for several more specific purposes. We will consider three such purposes commonly put forward for both size and levels tests: to aid in materials selection in educational programs; to track vocabulary growth, either for research or program evaluation; and to assist in identifying appropriate vocabulary learning goals.

MATERIALS SELECTION

Makers of size and levels tests have noted that such instruments could be helpful in selecting appropriate materials for language learners (Nation & Beglar, 2007; Schmitt et al., 2001; Stoeckel & Bennett, 2015). A useful paradigm for considering whether existing tests can adequately perform this function is Nation's (2007) four strands. In this framework, effective language programs are described as giving approximately equal attention to fluency development, meaning-focused input, meaning-focused output, and language-focused learning. The fluency development strand aims to help learners become better at using elements of the language that are already known. Therefore, in principle no new vocabulary should appear in materials for this strand. In meaning-focused input, learners' attention is focused on receiving and understanding a message rather than on learning new elements of the language, so reading passages in this strand should be

written at a level that can be understood independently. Nation (2007) has stated that learners should know at least 95% and preferably 98–99% of the running words in written materials for this strand. The optimal level of coverage for language-focused learning is less clear. Hu and Nation (2000) observed that there was no meaningful comprehension when learners knew 80% of the words in a text, suggesting that this level of coverage would be too low even when the focus of instruction is on the language. Schmitt et al., (2011) speculated that 85–95% coverage might be appropriate for this kind of learning.

Research indicates that differences between these levels of coverage are indeed significant. When learners know 98–99% of the words in a text, vocabulary knowledge is probably not a limiting factor in comprehension (Hu & Nation, 2000; Schmitt et al., 2011); however, the level of comprehension decreases rapidly with each successive 1% loss in coverage (Schmitt et al., 2011). Thus, if vocabulary tests are to assist in selecting materials for these different purposes, they need to be capable of distinguishing between whether learners know (a) all the words in a set of materials, (b) 95–99% of the words, (c) somewhere in the range of 85–95%, or (d) less than that. If a test is not accurate enough to do this, it is of limited use in materials selection.

TRACKING VOCABULARY GROWTH

A second commonly cited purpose of size and levels tests is to measure lexical growth for research or for assessing program efficacy (Nation & Beglar, 2007; Stoeckel & Bennett, 2015; Webb et al., 2017). For a test to serve this function, it must be able to establish an accurate baseline and be sensitive enough to track meaningful changes in lexical knowledge over time. For overall vocabulary size, Milton (2009) has summarized research showing that among L2 learners of English, the rate of growth tends to be approximately four words per classroom hour, with specific programs achieving gains from 300 to 500 words per year. This implies that tests need to be accurate enough to detect changes of this size. For high-frequency words, studies have estimated annual growth rates from 18 to more than 250 words (Agustin-Llach & Canga Alonso, 2016; Stoeckel, 2018; Webb & Chang, 2012), suggesting that levels tests need at least that degree of precision. This comes with the caveat that estimates of vocabulary growth in much of this research were made using instruments that have the very limitations that this article describes.

GOAL SETTING

A third commonly cited purpose for size and levels tests is to help learners set goals for their vocabulary learning (Nation, 2012b; Stoeckel & Bennett, 2015; Webb et al., 2017). Considering the previously discussed importance of knowing at least 98% of the words in a text, high-frequency vocabulary—defined here as the first 3,000 words of the lexicon (Schmitt & Schmitt, 2014)—is essential. This relatively small number of ubiquitous items provides a very large proportion of coverage of any natural text. Nation (2006b), for example, found that the first 1,000-word families provided 77.9% coverage of a wide range of written text types, with the second and third 1,000-word bands adding 8.2 and 3.7%, respectively. If tests are to help learners establish meaningful learning goals, they should be able to identify which high-frequency words, if any, are unknown. For learners who have mastered the first 3,000 words, rather than using frequency as a guide for further

learning, it may often be more efficient to focus on specialized or technical vocabulary depending on specific learning needs (Nation, 2013). Size tests do not separately assess knowledge of specialized vocabulary, but some levels tests include an academic vocabulary level for learners who need English for higher education.

In sum, when we consider some of the specific purposes for which size and levels tests have been developed, we get a better idea of how accurate these instruments need to be. Size tests ought to be able to detect changes in overall lexical knowledge at least as small as 500 words, preferably 200 words or less, if they are to be used to track annual lexical growth. Presumably, greater precision would be needed for shorter periods. Levels tests should be capable of (a) distinguishing between learners who have small differences in coverage of specified texts, namely four levels defined by 100%, 95–99%, 85–95%, and below 85% coverage; (b) detecting annual changes in knowledge of specific word bands at least as small as 250 words; and (c) identifying which words within a high-frequency or specialized word level are unknown so that they can be targeted for study.

TEST DESCRIPTIONS

Since the development of the Vocabulary Levels Test (VLT; Nation, 1983) in the early 1980s, several instruments have been made to measure written receptive vocabulary knowledge. To contextualize our later discussion of sources of inaccuracy, this section describes four such tests, chosen for the variety of features they represent. Due to space limitations, some other tests of vocabulary for reading, such as the New Vocabulary Levels Test (McLean & Kramer, 2015), could not be included.

VOCABULARY LEVELS TEST

The VLT was introduced by Nation in 1983 as a way to determine whether learners had gained mastery of high-, mid-, and low-frequency words as well as words that are common in academic discourse. Schmitt, Schmitt, and Clapham revised the VLT in 2001, creating two parallel forms of the test and conducting a preliminary validation study. This is the version most commonly in use today and the one described in the text that follows.

The VLT uses a matching format. Test items, referred to as "clusters," each contain six words and three definitions (Figure 1). The test has five levels, each with 10 clusters assessing 30 target words sampled from family-based lists (the distinction between word families, flemmas, and lemmas is discussed in the section "Word Grouping Principle").

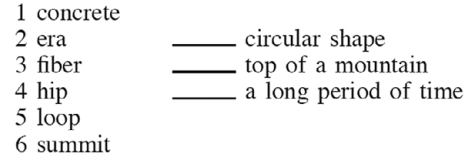

FIGURE 1. An example cluster from the Vocabulary Levels Test. From Schmitt et al. (2001).

The first level assesses knowledge of the combined first and second 1,000-word bands, and this is followed by levels that test the third, fifth, and tenth 1,000-word bands and one assessing knowledge of academic vocabulary drawn from Coxhead's (2000) Academic Word List. Within each level, stratified sampling was used to ensure that 15 nouns, 9 verbs, and 6 adjectives were assessed. Two parallel forms of the test were developed, and the target words in each level of each form were selected to represent the average item facility of that level from a bank of items that was trialed in the first stage of test development. Regarding score interpretation, scores of 26 or higher (out of 30) were put forth to represent mastery of a test level. However, no theoretical or evidence-based justification for this was given.

VOCABULARY SIZE TEST

Whereas the VLT was designed to determine whether learners had *mastery* of certain word bands, the Vocabulary Size Test (VST; Nation & Beglar, 2007) was developed to estimate overall written receptive vocabulary *size*. Stratified random sampling was used to select target words from each of the first fourteen 1,000-word frequency levels of family-based lists. As determining mastery of any given frequency level was not a goal of the test, the number of target words from each word band—just 10—was considerably smaller than in the VLT. The VST employs a multiple-choice format in which each item presents the target word followed by an example sentence which uses the word in a nondefining context. The four options include a definition of the target word plus definitions of three other words of similar frequency (Figure 2).

Nation later produced a version of the VST that includes five items for each of the first twenty 1,000-word frequency bands, which allowed for research with native speakers at secondary schools in New Zealand without having a ceiling effect (Coxhead et al., 2015). In addition to these two versions of the test, other variants have been produced including several bilingual forms in which answer choices are rendered in examinees' first language (Elgort, 2012; Karami, 2012; McLean et al., 2016; Nguyen & Nation, 2011; Zhao & Ji, 2018), a version with an added "I don't know" answer choice for each item (Zhang, 2013), and a version using a serial-multiple-choice item format (described in Item Format; Stoeckel & Sukigara, 2018). Regarding score interpretation, for versions of the test that are based on a target word sampling rate of 1:100, raw scores are multiplied by 100 to estimate overall vocabulary size. Similarly, raw scores are multiplied by 200 for the versions which use a 1:200 sampling rate.

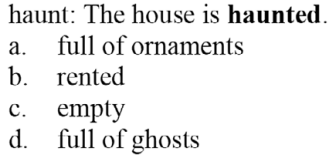

FIGURE 2. An example item from the monolingual 14,000-word version of the Vocabulary Size Test (Nation & Beglar, 2007) available at https://www.wgtn.ac.nz/lals/about/staff/paul-nation.

NEW GENERAL SERVICE LIST TEST

Stoeckel and Bennett introduced the New General Service List Test (NGSLT) in 2015. This instrument assesses knowledge of only high-frequency vocabulary, defined as the 2,801 entries in Browne et al.'s (2013) flemma-based New General Service List. Each of the five levels represents a 560-word frequency-based band. Approximately 40 words were randomly sampled from each band, data were collected to determine item difficulties, and then 20 items were selected for each level to represent the average and range of difficulty of the items in the level. The NGSLT employs a multiple-choice format like that used in the VST. In addition to the original monolingual format, there is a Japanese–English bilingual version (Stoeckel et al., 2018b). The authors of the NGSLT did not describe score interpretations in terms of mastery but instead stated that scores below 17 or 18 (out of 20) are meant to flag the level for intentional vocabulary study.

UPDATED VOCABULARY LEVELS TEST

Webb et al. (2017) made an updated Vocabulary Levels Test (UVLT). This is similar to Schmitt et al.'s (2001) VLT in that it has 30 target words per level (following a 15:9:6 ratio of nouns, verbs, and adjectives), words are sampled from word family lists, and it uses a matching format with three definitions and six words per cluster. Reflecting a shift in focus toward mastery of higher frequency words first, the UVLT differs from the VLT in that it assesses knowledge of words sampled from each of the first five 1,000-word frequency levels rather than from nonconsecutive levels or from a list of academic vocabulary. Additionally, words were selected from more up-to-date British National Corpus/Corpus of Contemporary American English frequency lists (BNC/COCA; Nation, 2012a) to reflect current English. The mastery criterion is 29 (out of 30) for the first three levels and 24 for the fourth and fifth levels. This difference was intended to reflect the greater relative importance of the higher frequency words in terms of the coverage they provide.

ISSUES IN VOCABULARY TESTING

We now examine three issues that can influence the accuracy of size and levels test scores. These are the item format, the word-grouping principles that are used in the lists from which words are sampled, and the size of the sample of target words.

ITEM FORMAT

In tests of written receptive vocabulary knowledge, two broad categories for item format are meaning-recall and meaning-recognition.

Meaning-Recall

Meaning-recall formats are used primarily in research. Test takers are shown a target word, often in the context of a nondefining example sentence, and asked to recall the word's meaning from memory without the assistance of a list of choices. There are two

kinds of meaning-recall tests. In *interviews*, test takers are asked to orally explain the meaning of target words or to directly translate them into the L1. An advantage of this method is it allows test administrators to ask unplanned follow-up questions when it is difficult to determine whether a word is known from an initial response (Schmitt, 2010). The second type, *written* meaning-recall tests, are like interviews except that responses are given in writing in the L1, either as paraphrases or direct translations of the target words. In comparison to interviews, this format is more convenient because it can be used with large groups of learners, and when tests are administered online, answers can be quickly compiled for marking. Both spoken and written meaning-recall formats are considered good, accurate ways to measure written receptive vocabulary knowledge (Kremmel & Schmitt, 2016; Nation & Webb, 2011) and are commonly used as criterion measures in validation studies of meaning-recognition tests (e.g., Kremmel & Schmitt, 2016; Stoeckel et al., 2019; Stoeckel & Sukigara, 2018; Webb et al., 2017; Zhang, 2013).

Meaning-Recognition

In meaning-recognition formats, learners select each target word's definition or L1 translation from a list of choices. In size and levels tests, these formats have included the previously described multiple-choice and matching item types as well as a newer serial-multiple-choice format (SMC; Stoeckel & Sukigara, 2018), which has been employed with the VST. The SMC format was designed specifically to reduce the difference in scores between meaning-recall and meaning-recognition formats. It uses computer administration to present answer choices one by one and does not permit examinees to revisit an option once they have rejected it as the correct response. Also, to prevent test takers from strategically choosing the final option if all others have been rejected, the number of options, which varies, is unknown to test takers.

Selected-response formats such as those used in meaning-recognition tests are considered "the best choice for test developers interested in efficient, effective measurement of cognitive achievement or ability" (Downing, 2011, p. 287) because they allow for wide sampling of the tested domain, are quickly and objectively scored, and, when well written, produce high estimates of reliability. Despite these advantages, meaning-recognition formats have important limitations in written receptive vocabulary testing.

Score Interpretation

The fundamental shortcoming relates to score interpretation. Correct answers are assumed to represent known words for the purpose of reading, and raw scores are interpreted in terms of the actual number of words that are known or mastery of the words in a test level. Between meaning-recall and meaning-recognition tasks, however, it has long been argued that the former is more similar to the aspect of lexical knowledge needed in reading because both meaning-recall and reading require word meaning to be evoked from memory without reference to a list of choices (Kremmel & Schmitt, 2016; Nation & Webb, 2011; Stewart, 2014; Stoeckel et al., 2019). This view is empirically supported by McLean et al. (2020), who found that meaning-recall test scores are significantly better predictors of reading proficiency than meaning-recognition scores.

There are two principal factors that make meaning-recognition item formats dissimilar from the lexical knowledge needed in reading. First, meaning-recognition requires a lower threshold of lexical knowledge than meaning-recall (Laufer & Goldstein, 2004; Nation, 2013; Nation & Webb, 2011; Schmitt, 2010), so words whose meanings are recognized from a list of choices may not be known when reading. Second, meaning-recognition items can be correctly answered through random guessing or the use of construct-irrelevant test strategies (Gyllstad et al., 2015; Kremmel & Schmitt, 2016; McLean et al., 2015; Stewart, 2014; Stewart et al., 2017; Stoeckel & Sukigara, 2018). These factors lead to overestimation of lexical knowledge for reading.

Although numerous studies have compared meaning-recall and meaning-recognition formats to ascertain the extent to which scores differ, some have potential sources of inaccuracy. In some cases, learners were instructed to skip unknown items or use an explicit "I don't know" option (e.g., Zhang, 2013), but these conventions are sometimes used differently (Bennett & Stoeckel, 2012; Zhang, 2013), which results in different scores for learners of similar ability (Stoeckel et al., 2016; Stoeckel et al., 2019). In several other studies, the meaning-recognition task was administered first (e.g., Gyllstad et al., 2015; Kremmel & Schmitt, 2016). This is useful for retrospective insight into test-taking behavior, but it exposes learners to the word form and meaning together, which can inform performance on the subsequent recall task. Finally, meaning-recall responses have been judged by comparing them to a predetermined bank of potential correct L1 translations (Aviad-Levitzky et al., 2019); this approach has real potential (as discussed in the following text), but incorrect responses (i.e., those not listed in the bank of correct answers) may need to be validated by comparison to human judgments, as there can be many ways to express the same concept.

We are aware of three studies with a total of five learner groups in which the recall test was administered first, examinees were required to answer every item, and each recall response received human scrutiny (Gyllstad et al., 2019; Stoeckel et al., 2019; Stoeckel & Sukigara, 2018). In these studies, the score difference between meaning-recognition and meaning recall was as if examinees had correctly answered 13.9–62.7% of the meaning-

TABLE 1. Meaning-recall and meaning-recognition score differences when the recall test is administered first, a response is required for each item, and all responses receive human judgment

				Scores		
Study	Item Type	Language of Options	*k*	Recall	Recognition	Overestimation[a]
Stoeckel & Sukigara, 2018	SMC	L1	80	31.7	38.4	13.9%
Stoeckel et al., 2019	MC	English	80	30.9	44.8	28.3%
Stoeckel et al., 2019	MC	L1	80	30.6	51.3	41.9%
Stoeckel & Sukigara, 2018	MC	L1	80	30.9	54.2	47.5%
Gyllstad et al., 2019	MC	L1	1,000	423.5	785.1	62.7%

Note: SMC = serial multiple choice; MC = multiple choice.

[a]Calculated as (correct recognition responses—correct recall responses)/incorrect recall responses (for explanation, see note 1).

recognition items that assessed words that were unknown at the meaning-recall level of knowledge (Table 1).[1] Both item type and test language were factors that affected score differences. The SMC format and monolingual options produced mean scores that were more similar to a criterion meaning-recall measure than traditional multiple-choice and bilingual options, respectively. Although other factors also affect the relationship between meaning-recall and meaning-recognition scores (e.g., learner proficiency relative to the difficulty of the tested words, Kremmel & Schmitt, 2016; Stewart & White, 2011), the range of values in the final column of Table 1 are useful for broadly understanding the potential extent of overestimated meaning-recall knowledge in size and levels tests.

On the basis of this previous research, what impact does item format have on estimates of vocabulary size? Taking the VST as an example, Table 2 displays estimated vocabulary sizes under meaning-recall and meaning-recognition formats at levels of meaning-recall knowledge from 1,000 to 8,000 words for three tests of the same length as those used in Beglar's (2010) validation of the VST: 40, 80, and 140 items to assess knowledge of the first 4,000, 8,000, and 14,000 words, respectively. The differences between recall and recognition scores shown in Table 2 reflect the highest and lowest proportion of unknown items answered correctly as reported in Table 1 (13.9% for a bilingual SMC format and 62.7% for a bilingual multiple-choice format) as well as a rate of 25%, which would be expected due to blind guessing under a four-option multiple-choice format. For example, the first row of data shows that meaning-recall performance indicating knowledge of 1,000 words is estimated to correspond with knowledge of 1,417 words under the SMC format (1,000 + 417 [i.e., 13.9% of the 3,000 words not known to the level of meaning-

TABLE 2. Differences in estimated vocabulary size at three levels of meaning-recognition knowledge for unknown items at the meaning-recall level

	Estimated Vocabulary Size			
		Meaning-Recognition		
		% of Unknown Items Answered Correctly		
Test Version	Meaning-Recall	13.9% (SMC)	25% (blind guessing)	62.7% (MC)
VST 1–4K (40 Items)				
	1,000	1,417	1,750	2,881
	2,000	2,278	2,500	3,254
	3,000	3,139	3,250	3,627
VST 1–8K (80 Items)				
	1,000	1,973	2,750	5,389
	2,000	2,834	3,500	5,762
	4,000	4,556	5,000	6,508
	6,000	6,278	6,500	7,254
VST 1–14K (140 Items)				
	1,000	2,807	4,250	9,151
	2,000	3,668	5,000	9,524
	4,000	5,390	6,500	10,270
	6,000	7,112	8,000	11,016
	8,000	8,834	9,500	11,762

Note: SMC = serial multiple choice; MC = multiple choice.

recall]), 1,750 words due to blind guessing of items not known to the level of meaning-recall under a four-option multiple-choice format, and 2,881 words when test strategies are also factored in with a multiple-choice format. Because there is less room for overestimation of meaning-recall knowledge as vocabulary size approaches the test ceiling, Table 2 shows that recall and recognition scores become more similar as they increase. Nonetheless, the difference between these two modalities can be quite large, often exceeding a thousand words.

For levels tests, the problems associated with item format are smaller, but not insignificant. Whereas size test scores are interpreted in terms of the actual number of words a learner knows even when many items are unknown, levels test are usually interpreted as to whether individual levels are mastered, with a high proportion of correct responses needed for mastery. As with size tests, when meaning-recall knowledge approaches the test ceiling, there is little room for overestimation. However, it has been theoretically posited (Stewart & White, 2011) and practically observed (Kremmel & Schmitt, 2016) that the use of test strategies can result in a higher proportion of unknown items being answered correctly when a learner's knowledge approaches level mastery. With a four-option multiple-choice format, we might therefore expect a level of success on unknown items similar to that shown in the rightmost column of Table 2. If so, roughly 60% of learners whose meaning-recall knowledge is one item below the mastery threshold would be able to achieve a mastery score, and about 36% whose recall knowledge is two items below this threshold could also do so. Using the four-option NGSLT as an example, even if the level mastery threshold were set at 20 out of 20 (a 100% success rate), many learners with actual knowledge of 18 or 19 items, which corresponds with 90% and 95% success rates, respectively, would be identified as having achieved mastery. Considering the substantial differences in comprehension that occur with small changes in coverage, the use of meaning-recognition formats with levels tests is not without consequences, even when the mastery threshold is set very high.

WORD GROUPING PRINCIPLE

Another characteristic of size and levels tests that can impact the accuracy of estimated employable vocabulary knowledge is the principle by which related word forms are grouped. In test development, target words are normally sampled from corpus-derived frequency-based wordlists. Such lists typically group word forms into lemmas, flemmas, or word families. A lemma consists of a headword together with inflected forms that are the same part of speech (POS) as the headword. For instance, the lemma for the verb *center* includes that headword and verbal uses of the inflected forms *centers*, *centered*, and *centering*. A flemma is similar to a lemma except the flemma takes no account of POS. That is, any use of the form *center* is included together with any use of the forms *centers*, *centered*, and *centering*. A word family, as used in L2 English vocabulary tests, generally refers to a headword together with all its inflectional and derivational forms up through Bauer and Nation's (1993) level 6 affix criteria (hereinafter WF6). The WF6 for *center* in Nation's (2012a) BNC/COCA lists includes the flemma constituents above as well as *centrist*, *centrists*, *central*, *centralization*, *centralize*, *centralized*, *centralizes*, *centralizing*, *centralism*, *centralist*, *centralists*, *centralities*, *centrality*, *centrally*, *centeredness*, and *centric* plus alternative spellings for 10 of these words.

In existing L2 vocabulary tests, learners are assessed on just one member of the lemma, flemma, or WF6 (usually the base form), and a correct response is assumed to indicate knowledge of the entire group of related word forms. When WF6 is used, this approach overestimates vocabulary knowledge, as all current L2 English research indicates that for EFL learners of a range of proficiencies, receptive knowledge of a headword is a poor indicator of knowledge of related derivational forms (Brown, 2013; McLean, 2018; Mochizuki & Aizawa, 2000; Schmitt & Meara, 1997; Ward & Chuenjundaeng, 2009). Overestimation is less extensive for flemma-based instruments because correct responses do not assume knowledge of derivational forms. However, there is evidence that receptive understanding of a word form in one POS does not always correspond with knowledge of the same word form in another POS (Stoeckel et al., 2018a), suggesting that flemma-based tests can also overestimate employable word knowledge.

What does the existing literature tell us about the *amount* of overestimation when tests assume a WF6 level of knowledge? To explore this question, Brown (2018) estimated the proportion of coverage of the complete set of texts in the BNC that is provided at each 1,000-word level of Nation's (2006a) BNC word lists by different levels of affixation knowledge and knowledge of irregular forms and alternative spellings. Summarizing Brown's findings for the first 1,000 WF6s, Table 3 shows that base forms alone provide about 67% of the total coverage that the 1K level offers. When irregular forms, alternative spellings, and compound forms (i.e., the "other" forms in Table 3) are added along with Level 2 affixes, this increases to 87.9%. These forms basically correspond with the flemma. Levels 3 through 6 of Bauer and Nation's (1993) framework provide the remaining 12.1% of coverage at the 1K level. This represents WF6 level of coverage.

Brown's analysis enables us to estimate the amount of employable coverage that a learner possesses relative to their test score, depending upon their level of affix knowledge. For instance, if a learner with a score of 98% on the 1K level of a WF6-based test has receptive knowledge of affixation up through only level 2 forms, the 1K test score (.98) could be multiplied by the coverage provided up through level 2 forms at the 1K level (.879, Table 3) to estimate that the employable coverage of 1K words for this learner is approximately 79.1%. This is substantially lower than the raw test score.

Although this is a good starting point for quantifying the extent to which use of WF6 overestimates employable coverage, real-world learners do not have all-or-nothing understanding of the affixes in each of Bauer and Nation's (1993) levels. Additionally, demonstrated knowledge of one word using a given affix is no guarantee that another word using the same affix will be known when the base word is known (McLean, 2018;

TABLE 3. Estimated proportion of coverage of the texts in the British National Corpus provided by knowledge at different levels of affixation for the first 1,000 words of Nation's BNC Word Lists (adapted from Brown, 2018)

Word Forms	Proportion of Coverage	Cumulative
Base word only	.670	.670
Other forms[a]	.042	.712
Level 2 forms	.167	.879
Level 3–6 forms	.121	1.000

[a]Includes irregular noun and verb forms, abbreviated forms, alternative spellings, and compound forms.

Mochizuki & Aizawa, 2000; Ward & Chuenjundaeng, 2009). This is because the known word may have been learned independently rather than understood as a base word plus affix. Data from two recent studies provide a more nuanced account of actual employable knowledge for learners of typical proficiency levels in one EFL context, Japanese universities.

In the first study, McLean (2018) examined knowledge of 12 base forms and 87 inflectional and derivational forms among 279 Japanese university students ranging from beginning to advanced proficiency. When learners knew a base form, they understood an extremely high proportion (.968) of related level 2 (inflected) forms but not so (.544) for levels 3 through 6 (derivational) forms (Table 4). This study also found very good comprehension (.980) for one irregular verb form, representing Brown's "other" category.[2]

In the second study, Stoeckel et al. (2018a) found that Japanese university students ($N = 64$) who understood a high-frequency word form in one POS were able to derive the correct meaning from the same word form when it occurred in another POS just 56% of the time. For example, learners who understood the use of *extra* as an adjective were often unable to understand its use as a noun. If the purpose of a written receptive vocabulary test is to ascertain when a word form can be understood in actual receptive use of the language (Schmitt et al., 2019), then this research suggests that knowledge of a base form in one POS should not be assumed from demonstrated knowledge of the same word form in another POS. Because many base forms from the first 5,000 WF6s have more than one POS (Table 5), this could substantially influence estimates of employable coverage.

Using Brown's (2018) BNC analysis together with the data from McLean (2018) and Stoeckel et al. (2018a), the employable coverage at each of the first five 1,000-word BNC

TABLE 4. Written receptive knowledge of irregular, inflected, and derived forms for known base words in McLean, 2018

Word Forms	Proportion Known When Base Form Was Known
Other form[a] (irregular)	.980
Level 2 Forms (inflections)	.968
Levels 3–6 Forms (derivations)	.544

[a]Based on just one irregular verb (*taught*).

TABLE 5. The proportion of content words with multiple parts of speech in each of the first five 1,000-word bands

BNC Band	Proportion of Content Words with Multiple POS
1	.544
2	.475
3	.438
4	.313
5	.231

Note: We calculated these proportions based on frequencies and POS designations for the British National Corpus from Lancaster University (http://ucrel.lancs.ac.uk/bncfreq/flists.html).

TABLE 6. Estimated employable coverage from base word knowledge on a WF6-based test of the first 1,000-word BNC Level for Japanese university students

	A	B	C
Word Forms	Proportion of Coverage Provided by Word Forms[a]	% Known by Japanese Learners	Employable Coverage (A * B)
Base word only	(.670)		
function words	.347	1.000	.347
content words (1 POS)	.147	1.000	.147
content words (2+POS)	.176	.560[b]	.099
Other forms	.042	.980[c]	.041
Level 2 forms	.167	.969[c]	.162
Level 3–6 forms	.121	.544[c]	.066
Total	**1.000**		**.862**

[a]From Brown (2018) and from Table 5.
[b]Calculated from data published in Stoeckel et al. (2018a).
[c]Calculated from data published in McLean (2018).

TABLE 7. Estimated employable coverage from base form knowledge in a WF6-based test of the first five 1,000-word BNC Levels for Japanese university students

BNC Level	Coverage (%)
1	86.2
2	79.7
3	81.6
4	84.7
5	87.7

Note: Calculated based on data from Brown (2018), McLean et al. (2020), and Stoeckel et al., (2018a).

levels for Japanese university students can be estimated. An example for the 1K level is shown in Table 6. Column A shows the proportion of coverage offered by specified word forms as reported by Brown (2018), with the base form category separated into function words, content words that have one POS, and content words with multiple POS. Column B shows the estimated proportion of each category that is known when knowledge of the base form is known, as reported by McLean (2018) and Stoeckel et al. (2018a). Column C, the product of A and B, shows the estimated employable coverage for these learners when knowledge of a base form has been demonstrated in one POS in a WF6-based test.

In this analysis, function words and the base form of content words that have just one POS were assumed to be known. Therefore, the values for these categories in column B are 1.000, and the values in columns A and C are identical, showing no loss in employable coverage. For the base forms of content words with multiple POS and levels 3–6 derived forms, however, there is substantial lost employable coverage. The final row of Table 6 shows that when interpreting test results of the 1K level, the Japanese university learners

in these two studies, on average, possessed only about 86% of the coverage indicated by test scores. The employable coverage for each of the first five 1,000-word bands of the BNC, calculated in the same way, is summarized in Table 7.

From Table 7, we can estimate the usable word knowledge at each 1,000-word level from the scores in a WF6-based levels test for Japanese university students. Take, for example, a hypothetical student with scores of 99, 98, 90, 80, and 70% on a test of the first five 1,000-word bands. These scores suggest very good knowledge of the first 2,000 WF6s. However, when these raw scores are multiplied by the corresponding values in Table 7 to account for lost employable coverage due to incomplete WF6 knowledge, a vastly different picture emerges (Table 8). The estimated employable coverage for the first two 1,000-word bands is just 85.1 and 78.4%, respectively. If, for example, the raw scores were used for materials selection, reading passages intended for use in the meaning-focused input or fluency strands of a language program would be nearly unintelligible without some assistance from a dictionary or glossing.

This analysis was based upon two studies whose participants were mostly representative of university students in Japan (with a small number of learners in McLean's 2018 study who were of unusually high proficiency). As these were average scores, some such students would certainly possess greater employable knowledge of word family constituents, but others would possess less, and the extent of overestimation for any given learner is not readily knowable from a WF6-based test score. There is, perhaps surprisingly, little research comparing base word knowledge to knowledge of derived forms, which is at the heart of the assumption underlying WF6-based tests, but the other study we are aware of is consistent with this analysis. Ward and Chuenjundaeng (2009) investigated knowledge of 16 base words and 16 related derived forms with two groups of Thai university students, finding that when learners had knowledge of a base form, they had knowledge of related derived forms 59.2% (high group) and 39.6% (low group) of the time.[3] Other investigations of L2 learners' receptive knowledge of derivational morphology have also shown far from complete understanding (Mochizuki & Aizawa, 2000; Sasao & Webb, 2017; see also Gardner, 2007).

Laufer and Cobb (2019) recently argued that WF6 remains usable because most derivational word forms utilize only a small number of common affixes, but this position relies on learners knowing such derivational forms. Unfortunately, when data from previous research is reanalyzed to include only words with frequently occurring affixes, the results are much the same (Table 9). For instance, when participants in McLean (2018) knew a base form, they could demonstrate knowledge of related derivational forms that

TABLE 8. Adjusted coverage figures for a hypothetical Japanese university student

	A	B	C
BNC Level	Test Score	Proportion of Test Score That Is Known	Estimated Employable Coverage (A * B)
1	99	86	85.1
2	98	80	78.4
3	90	82	73.8
4	80	85	68.0
5	70	88	61.6

TABLE 9. Written receptive knowledge of derived forms using common affixes when the base form is known

	% Known		
	Ward and Chuenjundaeng, 2009		
Affix	High Group	Low Group	McLean, 2018
~ion	58.5	31.1	47.3
~al			54.1
~er	66.9	31.2	94.2
~y			
~ly			52.2
~ate			
in~			
re~			75.3
~ity	41.2	57.5	24.0
~ant			
Overall	**59.3**	**38.4**	**62.1**

Note: Includes the 10 most frequent affixes in the Morpholex database developed by Sánchez-Gutiérrez et al. (2018). To the best of our knowledge, no studies have examined the ~y, ~ate, in~, or ~ant affixes under the same conditions as Ward and Chuenjundaeng (2009) or McLean (2018).

used frequently occurring affixes just 62.1% of the time, which is just a few percentage points higher than their performance across all affixes. Though the precise amount of overestimated word knowledge probably varies with learner proficiency and proximity of the L1 to English, there is no evidence to assume that use of WF6-based tests would be unproblematic for score interpretations with many, if not most, L2 learners of English.

It is important to note, however, that in the studies cited in the preceding text, decontextualized test formats were used to assess word knowledge. Because context can support lexical comprehension (Nation, 2013), the extent to which a more naturalistic reading task might enable learners to work out the meaning of derivational forms for known base words is an area in need of research. This would further clarify the appropriateness of using WF6 with tests of vocabulary for the purpose of reading. We also stress that our focus is on assessment, not pedagogy. The organization of word forms according to WF6 can be valuable for deciding how to treat related word forms in the classroom. A wonderful example is Gardner and Davies's (2014) Academic Vocabulary List, which is organized into both lemma- and family-based lists, the latter useful for helping learners to see the morphological relationship among family members.

TARGET WORD SAMPLE SIZE

We now turn to the issue of target word sample size. In the development of vocabulary size and levels tests, target items are sampled from large sets of words, often 1,000-word frequency-based bands. To achieve high estimates of internal reliability and a strong correlation with a criterion measure, it appears that a target word sample size of 30 items is sufficient (Gyllstad et al., 2015; Schmitt, et al., 2001).

A separate question, however, is how well a sample represents the population of words from which it was drawn. Vocabulary knowledge as a tested construct differs from assessment in many other areas of learning. Lexical knowledge is not a skill that, once acquired, can be applied to understand unrelated new words. Knowledge of lexis is built word by word, differing for example from basic addition and subtraction, areas of learning for which demonstrated mastery in a well-made test can be interpreted as the ability to solve a universe of similar problems. Thus, there are limitations regarding the inferences that can be made from knowledge of a random sample taken from a larger set of words (Gyllstad et al., 2015). A high estimate of internal reliability on a size or levels test indicates that the instrument can reliably measure knowledge of the target words in the test. Likewise, a strong correlation between scores on a vocabulary test and a criterion measure of the same words indicates that there is a strong relationship between measures of only the sampled words. This is separate from whether test scores accurately represent knowledge of an entire word band. Such distinctions are important because if the small number of words that are assessed is significantly more (or less) likely to be known than the population of words from which they were sampled, results will systematically over- (or under-) estimate vocabulary knowledge even if estimates of internal reliability or correlations with a criterion measure are high.

Perhaps this issue has received little attention because of an implicit assumption that the items within a frequency-derived word band are of similar difficulty. If this were the case, it would not matter which words were selected; they would be equally representative of the entire band. The empirical evidence does not support this assumption, however. Though mean scores on large, frequency-based word bands decrease with frequency (Aizawa, 2006; Beglar, 2010; Brown, 2013; Milton, 2007), there is considerable variation in difficulty for individual words within frequency bands (Beglar, 2010; Bennett & Stoeckel, 2014). This calls into question the accuracy with which small samples represent the average difficulty of a large population of words.

Though vocabulary levels and size tests are commonly used to estimate the total number of words known (e.g., Webb & Chang, 2012) or to determine level mastery (e.g., Chang, 2012), confidence intervals (CI) for individual scores are rarely if ever reported, and the number of items required for desired levels of confidence is under-explored. Using meaning-recall data from all 1,000 words at the 3K frequency level, Gyllstad et al. (2019) found that for a test consisting of 20 randomly selected items, the 95% CI was as much as 400 (i.e., ±200) of the 1,000 words assessed. Thus, a learner who correctly answers 10 out of the 20 items (50%) might know anywhere between 300 and 700 items in the 1,000-word band. The maximum 95% CIs for 30, 50, and 100-item tests were 326, 248, and 172 words, respectively.

Confidence intervals can also be estimated based on sample size and the proportion of correct responses. Using this approach, we calculated Clopper-Pearson (Clopper & Pearson, 1934) CIs for a hypothetical test with different target word sample sizes from a 1,000-word band. The Clopper-Pearson method is appropriate when the proportion of correct responses approaches or reaches 1 (McCracken & Looney, 2017), which corresponds with common mastery criteria for levels tests. We calculated CIs for two scoring outcomes. The first is at a score of 50% because this is where the CI is widest, revealing the largest potential difference between test scores and actual knowledge. The second is

TABLE 10. Clopper-Pearson confidence intervals for the number of known words for specified scores and sample sizes from a 1,000-word band

Score	Estimated Number of Words Known from Score	95% CI			90% CI		
		Low	High	Range	Low	High	Range
Out of 20							
10	500	272	728	456	302	698	396
18	900	683	988	305	717	982	265
19	950	751	999	248	784	997	213
20	1,000	832	1,000	168	861	1,000	139
Out of 30							
15	500	313	687	374	339	661	322
28	933	779	992	213	805	988	183
29	967	828	999	171	851	998	147
30	1,000	884	1,000	116	905	1,000	95
Out of 50							
25	500	355	645	290	375	624	249
48	960	863	995	132	879	993	114
49	980	894	1,000	106	909	999	90
50	1,000	929	1,000	71	942	1,000	58
Out of 100							
50	500	398	602	204	414	586	174
98	980	930	998	60	938	996	58
99	990	946	1,000	54	953	999	46
100	1,000	964	1,000	36	971	1,000	29

Note: CIs were calculated at openepi.com/Proportion/Proportion.htm.

where scores approach and reach 100% because this is where mastery has been defined in levels tests, and it is where CIs are smallest.

The values in Table 10, consistent with Gyllstad et al. (2019), indicate that when levels tests are used to estimate the number of words that a learner knows in a level (rather than complete mastery), massive error is possible for scores around 50%. Although the CIs for perfect or near-perfect scores are narrower, they are unsatisfactory when we consider the importance of knowing 95 or 98–99% of the words in a text. The most items per level in any existing levels test of the form-meaning link is 30 in both the VLT and UVLT, with the strictest mastery criterion set at 29 for the 1K and 2K levels of the UVLT. The 95% CI for that score (828–999 words) or even for a perfect score (884–1,000 words) casts doubt on whether a mastery score consistently corresponds with knowledge of 95% or more of the words in the level. The values in Table 10 also suggest that the current approach to test construction, in which items are randomly sampled, may be untenable for the needed level of precision for many testing purposes, even when a 90% CI is used. When a learner achieves a perfect score on a test of 100 items sampled from a 1,000-word band, the 90% CI falls outside of the level of 98% coverage. It is hard to imagine how a multiple-level test with 100-plus items per level could be practically used in most educational settings. For existing instruments, combining items from multiple test forms and, when time is limited, administering fewer than the full complement of levels, is a good way to increase precision.

FUTURE DIRECTIONS IN WRITTEN RECEPTIVE VOCABULARY ASSESSMENT

We have described three characteristics of existing written receptive vocabulary tests that weaken their accuracy: the use of meaning-recognition item formats; the assumption that a correct response on an item assessing knowledge of a single word form corresponds with knowledge of all flemma or WF6 members; and the use of small, random samples of target words. For each, we have presented empirical evidence to demonstrate that the scale of lost precision can be substantive. On this basis, our view is that existing size and levels tests lack the needed precision to estimate the number of words that a learner knows, to determine mastery of specific word bands, or indeed to fulfill many of the purposes that developers of these tests have put forward. Currently, the scale of uncertainty is simply too large for test users to have confidence in such determinations.

Despite this, we believe that existing tests remain useful if used properly. Inasmuch as these instruments effectively separate learners based on lexical knowledge, they are valuable for placement decisions in pedagogy and for grouping learners according to vocabulary knowledge in research. They can also be used together with other sources of information to support pedagogical decisions. Vocabulary test scores can supplement learner feedback and a teacher's own knowledge of their students to aid in material selection or in the identification of vocabulary for explicit study, for instance. However, existing size and levels tests by themselves should probably not be used for anything beyond ranking or grouping learners according to their vocabulary knowledge.

Ideally, the next generation of tests of vocabulary knowledge for reading will tell us something about a learner's understanding of each item in a word population or band, something that could be achieved with item response theory (IRT). That is, if meaning-recall data were gathered to estimate item difficulty measures for all the hundreds or thousands of items in a word set, perhaps IRT could be used to develop computer-adaptive tests of a reasonable length that could accurately estimate vocabulary size or level mastery based upon the probability of a learner with a given test score knowing each word in the set.[4] Likewise, individualized study lists could be made to include only the words in a level that a learner is unlikely to know. Because word difficulty can differ for learners from different L1 groups (Chen & Henning, 1985; Sasaki, 1991; Stoeckel & Bennett, 2013; Webb et al., 2017), this may need to be done separately for learners from different L1 backgrounds. However, once completed, this would be quite a powerful approach to test development that would certainly be one way to address the issue of target word sampling.

Another way to improve upon target word sampling, but only for mastery decisions in levels tests, might be to assess knowledge of the most *difficult* words in a word band rather than words that reflect the average difficulty of the band. With this approach, a high score (e.g., 29 out of 30) would be a more likely indicator of mastery because if learners know the most difficult items, they are also likely to know the easier ones. However, test developers would need to be abundantly clear that such instruments could not be used to estimate the proportion of words known in a test level. An advantage of this approach is that it could be adopted immediately with existing instruments that already have large item banks with known item difficulties (e.g., the VLT, NGSLT, or UVLT).

Regarding item format, meaning-recognition item types persist because of their convenience. Perhaps here again IRT could be used with computer adaptive meaning-

recognition tests to generate precise person measures that could be linked to estimates of word difficulty as described in the previous paragraph. The feasibility of this is unclear, however, because even though scores on meaning-recall and meaning-recognition tests strongly correlate (Gyllstad et al., 2015; Stoeckel et al., 2019), they are separate constructs. Another promising method already in use is to automate the marking of meaning-recall tests by comparing online meaning-recall responses to a database containing numerous viable correct LI translations (Aviad-Levitzky et al., 2019; Kim & McLean, 2019). One challenge with such an approach may be in identifying the many L1 synonyms that could be considered correct responses. However, if this method were used for the purpose of estimating person ability under IRT, the target words would not need to be randomly sampled, but could instead be chosen from a bank of items with known psychometric properties. Thus, perhaps target words could be carefully selected based upon how thoroughly the possible correct responses could be identified.

To prevent the overestimation of word knowledge associated with the use of WF6- or flemma-based instruments, echoing Kremmel (2016), we recommend lemma-based tests. Although WF6-based tests may be appropriate for some L2 learners of English, lemma-based instruments would provide more accurate estimates of vocabulary knowledge for many students and would make research findings for learners of different levels of proficiency more comparable.

CONCLUSIONS

A recent publication by Schmitt et al. (2019) called for more rigorous methods in the development of vocabulary tests going forward. We have examined several sources of inaccuracy in tests of written receptive vocabulary knowledge, demonstrated that the level of imprecision is arguably too great for these tests to fulfill the purposes they were designed for, and discussed possible innovations and more rigorous methodology for future test construction to address these shortcomings. Computer adaptive test administration and automatic marking of meaning-recall responses are among several interesting possible improvements that could be made to existing instruments.

NOTES

[1]Some research has quantified the difference between meaning-recall and meaning-recognition scores by calculating the percentage of score increase under the meaning-recognition format. For example, Webb et al. (2017) reported that the VLT (meaning-recognition) yielded scores 1.61 times higher than meaning-recall. A limitation of this approach, however, is that it depends upon test takers' level of proficiency. That is, the difference between the two test scores will be smaller for learners of higher proficiency because the number of unknown items as a proportion of the total score is smaller.

To illustrate this point, take two learners of different abilities sitting a 100-item meaning-recall test followed by a meaning-recognition test of the same words. For the sake of argument, we will assume a success rate of 25% due to blind guessing for unknown words on the recognition test. Learner A gets a meaning-recall score of 60. They correctly guess 10 of the 40 unknown words on the meaning-recognition test, thereby increasing that score to 70, or 1.17 times higher than meaning-recall. Learner B, however, gets a meaning-recall score of just 20. Because they correctly guess 20 of the unknown 80 items on the meaning-recognition test, they increase that score to 40, or a full 2.0 times higher than the meaning-recall measure. In short, using this approach can conflate properties of the test with learner characteristics, making it difficult to compare studies. The present study took a different approach. We first calculated the number of items that were answered incorrectly on the meaning-recall test and then

estimated the percentage of those items that were answered correctly on the meaning-recognition test. Using this approach with the preceding example, learners A (10 out of 40) and B (20 out of 80) would each have an overestimation rate of 25%.

[2]Although knowledge of a single word form is clearly not representative of all Brown's "other" forms, we know of no other studies which have examined Japanese learners' written receptive knowledge of irregular forms in relation to knowledge of the corresponding base forms.

[3]These figures are derived from the data published in Ward and Chuenjundaeng (2009).

[4]We are unaware of any such tests that have been developed from meaning-recall data, but see Brown and Culligan (2008) for a good example of a test using this approach with data from yes/no tests.

REFERENCES

Agustin-Llach, M. P., & Canga Alonso, A. (2016). Vocabulary growth in young CLIL and traditional EFL learners: Evidence from research and implications for education. *International Journal of Applied Linguistics*, *26*, 211–227. https://doi.org/10.1111/ijal.12090.

Aizawa, K. (2006). Rethinking frequency markers for English–Japanese dictionaries. In M. Murata, K. Minamide, Y. Tono, & S. Ishikawa (Eds.), *English lexicography in Japan* (pp. 108–119). Taishukan Publishing Company.

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*, 345–368. https://doi.org/10.1080/15434303.2019.1649409.

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, *6*, 253–279. https://doi.org/10.1093/ijl/6.4.253.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *31*, 101–118. https://doi.org/10.1177/0265532209340194.

Bennett, P., & Stoeckel, T. (2012). Variations in format and willingness to skip items in a multiple-choice vocabulary test. *Vocabulary Education and Research Bulletin*, *1*, 2. https://jaltvocab.weebly.com/publications.html.

Bennett, P., & Stoeckel, T. (2014). Word frequency and frequency of loanwords as predictors of word difficulty. *Vocabulary Education and Research Bulletin*, *3*, 2–3. https://jaltvocab.weebly.com/publications.html.

Brown, D. (2013). Types of words identified as unknown by L2 learners when reading. *System*, *41*, 1043–1055. https://doi.org/10.1016/j.system.2013.10.013.

Brown, D. (2018). Examining the word family through word lists. *Vocabulary Learning and Instruction*, *7*, 51–65. http://vli-journal.org/wp/vli-v07-1-2187-2759/.

Browne, C., & Culligan, B. (2008). Combining technology and IRT testing to build student knowledge of high frequency vocabulary. *The JALT CALL Journal*, *4*, 3–16. https://journal.jaltcall.org/issues/jaltcall-4-2.

Browne, C., Culligan, B., & Phillips, J. (2013). *The new general service list.* http://www.newgeneralservicelist.org

Chang, A. C.-S. (2012). Improving reading rate activities for EFL students: Timed reading and repeated oral reading. *Reading in a Foreign Language*, *24*, 56–83. https://nflrc.hawaii.edu/rfl/April2012/articles/chang.pdf.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing*, *2*, 155–163. https://doi.org/10.1177/026553228500200204.

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*, 404–413. https://doi.org/10.1093/biomet/26.4.404.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213–238. https://doi.org/10.2307/3587951.

Coxhead, A., Nation, P., & Sim, D. (2015). Measuring the vocabulary size of native speakers of English in New Zealand secondary schools. *New Zealand Journal of Educational Studies*, *50*, 121–135. https://doi.org/10.1007/s40841-015-0002-3.

Downing, S. M. (2011). Selected response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287–302). Routledge.

Elgort, I. (2012). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing*, *30*, 253–272. https://doi.org/10.1177/0265532212459028.

Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics*, *28*, 241–265. https://doi.org/10.1093/applin/amm010.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, *35*, 305–327. https://doi.org/10.1093/applin/amt015.

Gyllstad, H., McLean, S., & Stewart, J. (2019, July). *Empirically investigating the adequacy of item sample sizes of vocabulary levels and vocabulary size tests: A bootstrapping approach.* Paper presented at the Vocab@Leuven Conference, Leuven, Belgium. https://vocabatleuven.wordpress.com/program/

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

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*, 403–430. https://nflrc.hawaii.edu/rfl/PastIssues/originalissues.html

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

Kim, Y., & McLean, S. (2019, October). *Online self-marking typing, speaking, listening, or reading vocabulary levels tests*. Paper presented at the 27th Korea TESOL International Conference, Seoul, Korea. https://koreatesol.org/content/conference-book-full-version-pdf.

Kremmel, B. (2016). Word families and frequency bands in vocabulary tests: Challenging conventions. *TESOL Quarterly*, *50*, 976–987. https://doi.org/10.1002/tesq.329.

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*, 377–392. https://doi.org/10.1080/15434303.2016.1237516.

Laufer, B., & Cobb, T. (2019). How much knowledge of derived words is needed for reading? *Applied Linguistics*. Advance online publication. https://doi.org/10.1093/applin/amz051.

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

McCracken, C. E., & Looney, S. W. (2017). On finding the upper confidence limit for a binomial proportion when zero successes are observed. *Journal of Biometrics & Biostatistics*, *8*, 338–343. https://www.omicsonline.org/open-access/on-finding-the-upper-confidence-limit-for-a-binomial-proportion-when-zero-successes-are-observed-2155-6180-1000338.pdf.

McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*, *39*, 823–845. https://doi.org/10.1093/applin/amw050.

McLean, S., & Kramer, B. (2015). The creation of a new vocabulary levels test. *Shiken*, *19*, 1–11. http://teval.jalt.org/node/33.

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*, 26–35. http://vli-journal.org/wp/vli-v04-2-2187-2759/.

McLean, S., Stewart, J., & Batty, A. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge. *Language Testing*. Advance online publication. https://doi.org/10.1177/0265532219898380

McLean, S., Ishii, T., Stoeckel, T., Bennett, P., & Matsumoto, Y. (2016). An edited version of the first eight 1,000-word frequency bands of the Japanese-English version of the Vocabulary Size Test. *The Language Teacher*, *40*, 3–7. http://jalt-publications.org/node/2/articles/5244-edited-version-firsteight-1000-word-frequency-bands-japanese-english-version-v.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*, 142–151. https://doi.org/10.1177/026553228700400202.

Milton, J. (2007). Lexical profiles, learning styles and construct validity of lexical size tests. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 45–58). Cambridge University Press.

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

Mochizuki, M., & Aizawa, K. (2000). An affix acquisition order for EFL learners: An exploratory study. *System*, *28*, 291–304. https://doi.org/10.1016/S0346-251X(00)00013-0.

Nation, I. S. P. (2006a). *BNC-based word lists.* Victoria University of Wellington. http://www.victoria.ac.nz/lals/about/staff/paul-nation

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

Nation, I. S. P. (2012a). *The BNC/COCA word family lists.* http://www.victoria.ac.nz/lals/about/staff/paul-nation

Nation, I. S. P. (2012b, August). *Measuring vocabulary size in an uncommonly taught language.* Paper presented at the International Conference on Language Proficiency Testing in the Less Commonly Taught Languages, Bangkok, Thailand. http://www.sti.chula.ac.th/files/conference%20file/doc/paul%20nation.pdf

Nation, I. S. P., & Webb, S. (2011). *Researching vocabulary.* Heinle-Cengage ELT.

Nation, P. (1983). Teaching and testing vocabulary. *Guidelines*, *5*, 12–25.

Nation, P. (2007). The four strands. *Innovation in Language Learning and Teaching*, *1*, 2–13. https://doi.org/10.2167/illt039.0.

Nation, P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

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

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

Sánchez-Gutiérrez, C. H., Mailhot, H., Deacon, S. H., & Wilson, M. A. (2018). MorphoLex: A derivational morphological database for 70,000 English words. *Behavior Research Methods*, *50*, 1568–1580. https://doi.org/10.3758/s13428-017-0981-8.

Sasaki, M. (1991). A comparison of two methods for detecting differential item functioning in an ESL placement test. *Language Testing*, *8*, 95–111. https://doi.org/10.1177/026553229100800201.

Sasao, Y., & Webb, S. (2017). The Word Part Levels Test. *Language Teaching Research*, *21*, 12–30. https://doi.org/10.1177/1362168815586083.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan. https://doi.org/10.1057/9780230293977.

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework. *Studies in Second Language Acquisition*, *19*, 17–36. https://doi.org/10.1017/S0272263197001022.

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

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*, 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x.

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*. Advance online publication. 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*, 55–88. https://doi.org/10.1177/026553220101800103.

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

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

Stewart, J., McLean, S., & Kramer, B. (2017). A response to Holster and Lake regarding guessing and the Rasch model. *Language Assessment Quarterly*, *14*, 69–74. https://doi.org/10.1080/15434303.2016.1262377.

Stoeckel, T. (2018). High-frequency and academic English vocabulary growth among first-year students at UNP. *Journal of International Studies and Regional Development*, *9*, 15–30.

Stoeckel, T., & Bennett, P. (2013). Sources of differential item functioning between Korean and Japanese examinees on a second-language vocabulary test. *Vocabulary Learning and Instruction*, *2*, 47–54. http://vli-journal.org/wp/vli-v02-1-2187-2759/.

Stoeckel, T., & Bennett, P. (2015). A test of the new General Service List. *Vocabulary Learning and Instruction*, *4*, 1–8. http://vli-journal.org/wp/vli-v04-1-2187-2759/.

Stoeckel, T., Ishii, T., & Bennett, P. (2018b). A Japanese–English bilingual version of the New General Service List Test. *JALT Journal*, *40*, 5–21. https://mail.jalt-publications.org/articles/24292-japanese-english-bilingual-version-new-general-service-list-test.

Stoeckel, T., & Sukigara, T. (2018). A serial multiple-choice format designed to reduce overestimation of meaning-recall knowledge on the Vocabulary Size Test. *TESOL Quarterly*, *52*, 1050–1062. https://doi.org/10.1002/tesq.429.

Stoeckel, T., Bennett, P., & McLean, S. (2016). Is "I don't know" a viable answer choice on the Vocabulary Size Test? *TESOL Quarterly*, *50*, 965–975. https://doi.org/10.1002/tesq.325.

Stoeckel, T., Ishii, T., & Bennett, P. (2018a). Is the lemma more appropriate than the flemma as a word counting unit? *Applied Linguistics*. Advance online publication. https://doi.org/10.1093/applin/amy059.

Stoeckel, T., Stewart, J., McLean, S., Ishii, T., Kramer, B., & Matsumoto, Y. (2019). The relationship of four variants of the Vocabulary Size Test to a criterion measure of meaning recall vocabulary knowledge. *System*. Advance online publication. https://doi.org/10.1016/j.system.2019.102161.

Ward, J., & Chuenjundaeng, J. (2009). Suffix knowledge: Acquisition and applications. *System*, *37*, 461–469. https://doi.org/10.1016/j.system.2009.01.004.

Webb, S., & Chang, A. C.-S. (2012). Second language vocabulary growth. *RELC Journal*, *43*, 113–126. https://doi.org/10.1177/0033688212439367.

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

Zhang, X. (2013). The I don't know option in the Vocabulary Size Test. *TESOL Quarterly*, *47*, 790–811. https://doi.org/10.1002/tesq.98.

Zhao, P., & Ji, X. (2018). Validation of the Mandarin version of the Vocabulary Size Test. *RELC Journal*, *49*, 308–321. https://doi.org/10.1177/0033688216639761.