Routledge
Taylor & Francis Group

Language Assessment Quarterly

ISSN: 1543-4303 (Print) 1543-4311 (Online) Journal homepage: http://www.tandfonline.com/loi/hlaq20

Interpreting Vocabulary Test Scores: What Do Various Item Formats Tell Us About Learners' Ability to Employ Words?

Benjamin Kremmel & Norbert Schmitt

To cite this article: Benjamin Kremmel & Norbert Schmitt (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

To link to this article: http://dx.doi.org/10.1080/15434303.2016.1237516

Published online: 15 Nov 2016.

LANGUAGE ASSESSMENT QUARTERLY
2016, VOL. 13, NO. 4, 377–392
http://dx.doi.org/10.1080/15434303.2016.1237516

Routledge
Taylor & Francis Group

Interpreting Vocabulary Test Scores: What Do Various Item Formats Tell Us About Learners' Ability to Employ Words?

Benjamin Kremmel[a,b] and Norbert Schmitt[a]

[a]University of Nottingham, Nottingham, United Kingdom of Great Britain and Northern Ireland; [b]University of Innsbruck, Innsbruck, Austria

ABSTRACT

The scores from vocabulary size tests have typically been interpreted as demonstrating that the target words are "known" or "learned." But "knowing" a word should entail the ability to use it in real language communication in one or more of the four skills. It should also entail deeper knowledge, such as knowing the word's derivative forms or collocations. However, little is known about how informative various vocabulary item formats are about lexical employability or mastery of aspects of word knowledge. That is, we have little idea about what score interpretations are warranted from various item formats. This article reports on two initial studies that investigated four form-meaning item formats (multiple matching, multiple-choice, and two types of cloze) to determine how informative they can be about lexical employability and knowledge of derivatives and collocations. Results showed that the four formats performed poorly and that it may not be warranted to interpret scores from these formats as showing that vocabulary is known to the level where it can be used when reading. Furthermore, the findings suggest that scores from these tests cannot be taken to mean that the words are known to any deeper degree than just the form-meaning link.

Introduction

Vocabulary size tests are widely used in research and pedagogy, with a wide variety of formats available. Size tests have been used for many purposes, ranging from a proxy for language proficiency in general, to a diagnostic to determine whether learners possess the lexical resources to be functional in the four skills. However, nearly all of the validation carried out on vocabulary size tests to date has focused on technical aspects, such as reliability and target word sampling from frequency lists. Little research has explored how the resulting size test scores can be interpreted. That is, we know little about what various item formats tell us about the underlying vocabulary knowledge of learners. This surprising gap concerns at least two facets. The first is what various item formats tell us about the learners' ability to use the target words in the four skills. Stated simply, if a learner answers an item correctly, can that learner actually use that target word, for example, understand it when reading? The second facet concerns the aspects of word knowledge that can be assumed to be known if an item is answered correctly. That is, how "deep" is the learners' knowledge of the target item (Schmitt, 2014)? For example, can we infer they know all of the members of the larger word family, the word's collocations, or how to use the word appropriately? These two facets are directly related to our ability to interpret vocabulary size test scores in any meaningful way, because a raw size score without knowing what the learner can do with that amount of vocabulary is of little use. This article reports on two initial studies that explore the informativeness of a variety of receptive and productive item formats.

CONTACT Benjamin Kremmel benjamin.kremmel@uibk.ac.at University of Nottingham, School of English, University Park, Nottingham NG7 2RD, United Kingdom.

Background

The choice of item format should correspond to the purpose of a test and depends on the kind of information a test developer wants to elicit. Nation and Webb (2011) state that "[w]hen designing a vocabulary test, careful thought needs to be given to the item type that is used to make sure that it is suited to the kind of knowledge it is supposed to measure" (p. 219). However, in vocabulary testing, decisions about which item format to use seem to be determined more by tradition and concerns for practicality than empirically grounded rationales and a close consideration of test purpose. The designs of even the most prominent vocabulary tests, such as the Vocabulary Levels Test (VLT) (Nation, 1990; Schmitt, Schmitt, & Clapham, 2001) and the Vocabulary Size Test (VST) (Nation & Beglar, 2007), appear to have been guided mainly by intuitions about feasibility, without fully accounting for what any particular format (and the scores it yields) can and cannot tell about the lexical abilities of a test taker.

For example, these two tests are usually considered tests of receptive vocabulary knowledge (i.e., vocabulary for reading and listening), with the VST explicitly placed as a test of reading vocabulary: "Users of the test need to be clear what the test is measuring and not measuring. It is measuring written receptive vocabulary knowledge, that is, the vocabulary knowledge required for reading" (Nation, 2012a, no page number). However, neither test has validation evidence to demonstrate that correctly answered target words can indeed be used either in reading or listening. In fact, there is good reason to believe that the multiple matching (VLT) and four-option multiple-choice (VST) formats the tests use are ill suited for this purpose. Fluent reading (and listening) requires quick recognition of the word form, and automatic recall and retrieval of the corresponding meaning, so that cognitive resources can be applied to meaning construction from the text (Grabe, 2009). Thus, vocabulary needs to be known to the *meaning recall* level (Schmitt, 2010) to reach lexical employability (i.e., make fluent reading possible). In a reading situation the authentic task for a learner to perform is to recall the meaning of the word form they are exposed to without any help or meaning options to choose from. But matching and multiple-choice items are *recognition* formats, where options are given and must be selected from. Such recognition formats are clearly incongruent with real-world reading, because no book provides multiple definitions to choose from for (unknown) words in the text (Nation & Webb, 2011).

Correctly answered items on vocabulary size tests are usually interpreted as being "known" or "learned," but the problem is in defining this. As suggested above, assumptions that being "known" relates to the ability to use words in one or more of the four skills is largely unsubstantiated for most vocabulary tests. (See the Listening Vocabulary Levels Test for a recent exception [McLean, Kramer, & Beglar, 2015].) Being "known" could also be interpreted as having mastery of the various aspects of word knowledge, such as collocation, register, and derivation (Nation, 2001). But little is known about how informative different item formats are about these word knowledge aspects. In fact, for most vocabulary tests, the level and type of knowledge the tests show is seldom specified. However, specifying this construct is generally regarded a prerequisite for score interpretations to be meaningful (e.g. Bachman, 1990; Bachman & Palmer, 2010).

In practice, most commonly used vocabulary formats (e.g., multiple-choice, matching, cloze, translation) address only the form-meaning link. That is, they either provide the word form and ask for recognition or recall of the meaning, or vice versa. (The Yes/No format is more difficult to interpret, and scores can vary greatly, depending on the formulas and pseudowords used to adjust for overestimation (e.g., Pellicer-Sánchez & Schmitt, 2012; Stubbe, 2012). However, even knowledge of the form-meaning link is not straightforward, because Laufer and Goldstein (2004) show that there is a hierarchy of the form-meaning relationship, with form recall being the most difficult and meaning recognition being the easiest (based on a translation format). So scores of form-meaning knowledge will depend on the type of form-meaning format being used.

Most form-meaning vocabulary tests use short, discrete, context-independent, selected item formats (Read, 2000). These formats, such as multiple-choice and matching, remain popular because of their apparent aptness for measuring individual words: they provide "objective" scores, are practical in test

administration and scoring, and allow for a relatively high sampling rate in a relatively short amount of time. However, these selected response formats are not without flaws. For example, Goodrich (1977) cautions that the nature of the distractors has a considerable impact on the measurement and its outcome. But perhaps the greatest criticism of these formats concerns their guessability. For example, Kamimoto (2008) and Webb (2008) suggest there is a 17% chance of learners blind guessing correct responses in the multiple matching format of the VLT. In their multiple guessing simulations on the VLT, Stewart and White (2011) found that candidates' scores are generally and consistently inflated by 16–17 points on a 99-item VLT test "until over 60% of words are known, at which point the score increase due to guessing gradually begins to diminish" (p. 378). In a more recent article Stewart (2014) argues that multiple-choice items indeed inflate scores on vocabulary size tests, advising that, whenever possible, this format should be avoided when measuring receptive vocabulary breadth. However, his argument is again based on simulations rather than real test-taker data.

It seems logical that to interpret the meaning of vocabulary size scores, the scores must be specified to either skill employability or word knowledge mastery, or both. However, to date, little validation research has focused on score interpretation, and so our understanding of the workings of vocabulary formats is currently insufficient. This is partly because few studies have explored the behaviour of the item formats themselves and what they can and cannot tell us. One exception is Paul, Stallman, and O'Rourke (1990). They compared multiple-choice and yes/no scores with scores from an interview that focused on how well the target words could be used in reading. They tested 20 high-ability and 20 low-ability readers on their knowledge of 44 polysemous words and found that both the multiple-choice format (between .66 and .82) and the yes/no format (between .69 and .81) correlated significantly and highly with the interview measure. They concluded that their four-option multiple-choice format was the most suitable for testing breadth of knowledge for reading, not least because it gave a representative indication of the knowledge students have of specific meanings of words. However, in line with many of the studies mentioned above, they also found the problematic influence of test-taking strategies and report that "guessing" was frequently used, particularly by lower ability students (21% of the cases), although this strategy was only successful in about a third of the attempts. Echoing earlier findings by Pike (1979), which showed that multiple-choice formats were among the most efficient item types, the Paul *et al.* study seems to suggest that the multiple-choice format might be suitable for testing vocabulary size connected with reading.

In one of the only studies to explore a larger numbers of formats, Henning (1991) explored eight different multiple-choice formats in a large-scale study into the functioning of TOEFL vocabulary items. Analyzing 190 test-takers' scores on a total of 1,040 items counterbalanced across eight format conditions, he found that items embedded in a reading text appeared to outperform the then-traditional TOEFL vocabulary item, in which the target was part of a lengthy sentence and needed to be matched with a synonym. However, none of the correlational differences in his study reached significance, rendering his claims relatively tenuous. In addition, his choice of using the vocabulary total scores on these experimental items as a criterion measure against which to correlate scores from different formats is questionable at best. In any case, Henning's main research interest was the effect of the degree of contextualisation of the different item versions, so the study is more an in-depth analysis of variations of one format (multiple-choice) than an investigation into how format scores can be properly interpreted.

The lack of research into vocabulary score interpretation is surprising, because this is an essential component for matching vocabulary tests to particular purposes. The present study attempts to address this gap in the research by investigating the following research questions concerning the employability and knowledge of target vocabulary.

RQ1. Can scores from four form-meaning item formats be interpreted as providing accurate information about L2 learners' ability to employ the target vocabulary in reading, as indicated by a concurrent meaning recall measure?

RQ2. Can scores from four form-meaning item formats be interpreted as providing accurate information about L2 learners' knowledge of the target vocabulary's derivatives, as indicated by a concurrent derivative measure?

RQ3. Can scores from four form-meaning item formats be interpreted as providing accurate information about L2 learners' knowledge of the target vocabulary's collocations, as indicated by a concurrent collocation measure?

Pilot study

The pilot study was designed to carry out an initial exploration of what various vocabulary item formats show about the ability of learners to employ target words in a skill, in this case reading.

Participants

The participants included 18 English native speakers (NS) and 12 non-native English speakers (NNS), all students at a school of English at a British university. The native speakers were all undergraduate students (16 female and 4 male) with a mean age of 18.9, whereas 10 of the non-native speakers were postgraduate students and two undergraduate students (8 female and 4 male) with an average age of 26.9. The nine different L1s of the non-native speakers included French, Italian, Bosnian, Portuguese, Lithuanian, Greek, Arabic, Chinese, and Dutch.

Target vocabulary

In preparation for the pilot study, a pretest was conducted with 15 NS and 10 NNS, all students at a British university, who were not part of the pilot study. They were administered words collated from the 11 K to 20 K levels of the two 20 K VST versions published online (Nation, 2014). To ensure that there would be a mixed range of target vocabulary, some of which would be known to parts of the target population, and some of which would not, 36 of these items were selected for their average facility values, ranging from .32 (*cerise*) to .88 (*headstrong*) across both groups, with a total average facility value of .61 across all items and groups.

Form-meaning item formats

Four item types that focused on various aspects of the form-meaning link were investigated: one targeting form recognition, one targeting meaning recognition, and two targeting form recall. For form recognition, we chose the item format from the most commonly used vocabulary size test, the Vocabulary Levels Test (VLT) (Schmitt *et al.*, 2001). It uses a multiple matching (MM) format, in which examinees must write the number of the word that matches the short definition (Figure 1).

For meaning recognition, we selected the item format from a relatively new test that is increasingly being used, the Vocabulary Size Test (VST) (Nation & Beglar, 2007). The VST uses a traditional four-option multiple-choice (MC) format (Figure 2).

The two form recall formats provided a definition of the target word, and examinees were required to write the word in a type of cloze design. The initial letter of the word and an indication of the number of letters of the target word were provided to disambiguate possible answers. An alternative form recall format used in the Computer-Adaptive Test of Size and Strength (CATSS) (Laufer, Elder, Hill, & Congdon, 2004) with no indication of the length of the target word (i.e., just a single blank provided) was considered but discarded as problematic before the pilot because it was prone to ambiguous candidate answers. The first form recall format provided only a short definition as the prompt (DEF) (Figure 3).

1. alimony	
2. beagle	___ small dog with long ears
3. counterclaim	
4. kestrel	___ statement made opposing a previous statement
5. proclivity	
6. reprise	___ money for the care of children, paid regularly after a divorce

Figure 1. Multiple matching (MM) VLT form recognition format.

beagle: He owns two **beagles**.
a) fast cars with roofs that fold down
b) large guns that can shoot many people quickly
c) small dogs with long ears
d) houses built at holiday places

Figure 2. Multiple-choice (MC) VST meaning recognition format.

a small dog with short legs and long ears, used in hunting
b __ __ __ __ __

Figure 3. Definition-only (DEF) form recall format.

a small dog with short legs and long ears, used in hunting
He owns a **b** __ __ __ __ __ .

Figure 4. Word in non-defining context (CON) form recall format.

Because context has been shown to facilitate some item formats (Martinez, 2011), the second form recall format also presented the target word in a non-defining short sentence context (CON) (Figure 4). Mirroring the comparison of the recognition formats, where the MC format provided such contextualisation, it was assumed that this might have a facilitative effect in this format.

Measure for employability in reading

We believe that vocabulary size scores are often interpreted in their employability in the four skills. To explore whether this is justified, we needed a measure of vocabulary employability in these skills. It was impractical to measure all four skills, but because a particularly close relationship has been shown between vocabulary and reading (e.g., Alderson, 2005), we decided to focus on this skill. We could find no research investigating whether correctly answered words on vocabulary tests can actually be comprehended when they occur in authentic reading texts. One reason for this lack of research is probably the difficulty in designing a study capturing this comprehensibility. An obvious approach is to first test target words and then embed these words in texts to see if they can be comprehended during reading. However, many factors can affect comprehensibility, including the richness of the context clues, background knowledge of the text topic, and the grammatical structures that words occur in (e.g., Grabe, 2009). These factors make this ecologically valid approach very difficult to realize in a workable research design.

We decided to operationalize employability in reading in a slightly less direct manner, by measuring the level of lexical mastery required to read. When reading, the word form is given on the page, the reader must recognize this form and quickly recall its meaning, not to slow down the fluent reading process. This is sometimes called *sight vocabulary* by reading scholars and in vocabulary knowledge relates to a

meaning recall level of mastery (Schmitt, 2010). We chose interviews as a method of tapping into this meaning recall knowledge. Although interviews are time-consuming, they "have the value of being a stringent unguided test of knowledge" (Nation & Webb, 2011, p. 216).

Procedure

The 36 target items were clustered into four groups of 9 items. These clusters were balanced according to pretest facility values. Four test versions were then compiled, featuring all four item clusters in all item type formats in a Latin Square design. In this way the target word *beagle*, for instance, was presented as a MM item in Version A, as an MC item in Version B, as a definition-only recall item (DEF) in Version C, and as a recall item with non-defining sentence context (CON) in Version D (Table 1).

One of the four test versions was administered individually as a paper-pencil test to each of the participants. After they had taken the test, each candidate was interviewed face-to-face by the first researcher. Candidates were asked to recall the meaning of the target words, probing which item type could provide the most informative picture of word knowledge and best represent the meaning recall knowledge of candidates. A relatively precise definition was required in the interview to facilitate the interpretation that the words could be comprehended in most reading contexts. While partial knowledge (e.g., knowing that a *beagle* is a type of dog) might be sufficient in some reading contexts, more precise knowledge (e.g., how a *beagle* looks) may be required in other contexts. Likewise, the form recall items (DEF and CON) required exact spelling. All measures were scored on a dichotomous correct or incorrect basis.

Results

Candidates' test scores were compared to their "verified" word meaning knowledge scores from the interview, following Schmitt *et al.* (2001). As illustrated in Figure 5, there are four possible outcomes when matching these dichotomously scored items. Cases match if the candidate was either awarded a point in both the vocabulary test item and the respective meaning recall measure (A) or if the candidate answered neither correctly (D). If a candidate was awarded a point for the vocabulary test item but did not show sufficient knowledge of the word in the interview, it is a case of overestimation (B) (i.e., the test score overestimating the meaning recall knowledge of a candidate). Vice versa, if a candidate was not

Table 1. Item cluster distribution across formats and test versions.

	Version A	Version B	Version C	Version D
Multiple Matching (MM)	1–9	28–36	19–27	10–18
Multiple Choice (MC)	10–18	1–9	28–36	19–27
Form Recall with Definition Only (DEF)	19–27	10–18	1–9	28–36
Form Recall with Definition and Context (CON)	28–36	19–27	10–18	1–9

		Interview meaning recall measure	
		Known	Not Known
Test item	Correct	Match (A)	Overestimation (B)
	Incorrect	Underestimation (C)	Match (D)

Figure 5. Contingency table of matching/mismatching results.

awarded a point for the test item but was judged to actually know the meaning of the word in the interview, the test item underestimated the candidate's meaning recall knowledge (C).

For the first analysis, the A and D cells were combined (matches), as were the B and C cells (mismatches), to obtain a global view of how well the various item formats indicated meaning recall knowledge for our low-frequency target words (Table 2). Overall, the multiple-choice format best represented the word knowledge of the candidates, with 83% matches of the 270 cases. This slightly outperformed the multiple matching format (80.7%), although the behaviour of the two formats is so similar as to essentially be considered the same. Conversely, the two recall formats performed more poorly, with only 70.4% matches (with sentence context) and 64.4% (definition only). This means that in about one-third of the cases, both form recall formats produced scores that did not match with the interview result. However, the addition of the non-defining sentence context did produce better results than the definition-only format.

Both overestimation and underestimation are measurement errors in that there is a mismatch between the test score and a participant's meaning recall knowledge (as verified by the interview). While both cases can be subsumed as mismatching cases and thus signify a problem with the measurement tool, they represent two very different item behaviours that warrant closer analysis. Our next analysis looked at the two behaviours separately (Table 3).

Similar to previous studies (e.g., Paul *et al.*, 1990; Stewart & White, 2011; Webb, 2008), our results indicate that the recognition formats have a problem with overestimation (18% of the cases in MM format and 12% of the cases in MC format). Conversely, the form recall formats unsurprisingly did not have overestimation problems (less than 1%), because filling in the blanks correctly through guessing or test-taking strategies if the target word is unknown is highly unlikely. More important, though, is the finding that the form recall formats considerably underestimated the word meaning recall knowledge of candidates (definition-only: 35%; sentence context: 29%). Although this tendency could be expected, its extent is surprising, with the gap in strength of knowledge between form recall and meaning recall even larger than that found by Laufer *et al.* (2004). It has been well documented that recognition item formats are susceptible to guessing and other factors inducing measurement error, but the literature has not previously identified the tendency toward such substantial underestimation for form recall formats. Overall, our pilot findings suggest that recognition and recall formats have very different characteristics concerning overestimation and underestimation of meaning recall vocabulary knowledge.

Note that despite pretesting and careful selection, the target words might have been relatively challenging for the participants. Even in the recognition formats, candidates did not arrive at the correct answer in either the test item or the criterion measure in 28% and 35% of the cases. In the more difficult recall measures, this value reaches 49% in both formats. Thus, our results might be very different with higher-frequency targets, and this has been taken into account in the main study.

Table 2. Correspondence between item formats and the interview meaning recall measure (overall percentage of cases).

	Form Recognition Multiple Matching	Meaning Recognition Multiple-Choice	Form Recall Definition-Only	Form Recall Sentence Context
Match	80.7%	83.0%	64.4%	70.4%
No match	19.3%	17.0%	35.6%	29.6%

Table 3. Correspondence between item formats and the interview meaning recall measure (percentage of cases).

	Form Recognition Multiple Matching	Meaning Recognition Multiple-Choice	Form Recall Definition-Only	Form Recall Sentence Context
Match (point)	53.0%	48.1%	15.6%	21.5%
Overestimation	18.1%	11.5%	0.7%	0.7%
Underestimation	1.1%	5.6%	34.8%	28.9%
Match (no point)	27.8%	34.8%	48.9%	48.9%

Likewise, the time-consuming nature of using interviews as a criterion measure meant that the number of participants was relatively small, limiting the generalizability of findings. Therefore, adjustments were made in the main study methodology to increase the number of participants.

Main study

In the main study we wished to further explore the relationship between the four item formats and target word employability, but with greater participant numbers. We also expanded the investigation to explore the relationship between the formats and enhanced knowledge of the target words, namely, their derivations and collocations.

Participants

For the main study, intermediate EFL learners were used as a target population because they are more likely to be takers of vocabulary size tests than native speakers or highly proficient non-native speakers. Data were gathered from 99 Austrian EFL learners in their penultimate and final years of secondary education. Only 10 participants indicated an L1 other than German. The mean age of the participants was 16.9 years; 56 were female and 31 male (12 did not indicate).

Target vocabulary

To select an appropriate target frequency level for the participants of the main study, a reduced version of the VLT (2 K, 3 K, and 5 K) was administered to 25 Austrian EFL learners in their penultimate year of secondary education. After reordering the targets according to their more up-to-date frequency information of the BNC-COCA word family list (Nation, 2012b), a combined frequency list based on data from the British National Corpus and the Corpus of Contemporary American English, the 3 K frequency band appeared to be the most appropriate level for target sampling, with a mean facility value of .72. Therefore, 36 words were sampled from the 3 K BNC-COCA lists (Nation & Webb, 2011). The words were selected for their part of speech to be able to cluster them together into groups of three for the multiple matching format. Each target also had to have at least three derivational forms, because we included a derivation test in our test battery (see below).

Item formats

Four item test formats were created for each target word. These were the same as used in the first study, because we wished to continue exploring the relationship between these formats and reading employability. In addition, we expanded the research design to include measures of two word knowledge aspects.

The first was derivation knowledge (DER). A test of derivative knowledge seemed particularly necessary, because most vocabulary tests (e.g., VLT and VST) are based on the counting unit of word families. However, this only works if learners actually know the various members of the families. This has been called into question, at least for productive mastery of derivatives (Schmitt & Zimmerman, 2002). Following the theme of vocabulary and reading, we considered the format that would best reflect the use of derivative knowledge in reading. Clearly, productive formats would not be congruent (e.g., the *Test of Derivatives*, Schmitt & Zimmerman, 2002). We reasoned that when readers come across a word family member (e.g., *darkness*), they must recognize or establish that it is related to the particular headword of the family, which they might know (e.g., *dark*). For each of the target words, three derivational forms (sampled from the BNC-COCA frequency lists) were given to the test takers, and they were asked to write down the headword on which these were based (Figure 6). Three derivational forms (rather than two or merely one) were provided to help learners arrive at the required headword.

Write down the word which you think is the basis of the words displayed. Do not just copy one of the three words.

inaccuracy
accurately
accuracies

Figure 6. Receptive derivative knowledge format (DER) (answer = *accurate*).

Collocation knowledge is now recognized as an important component of lexical knowledge (e.g., Barfield & Gyllstad, 2009; Nesselhauf, 2005), so it seemed a reasonable second word knowledge aspect to add to the test battery. We looked for a format that focused on multiple collocates for the target words, and Eyckmans's (2009) discriminating collocations (DISCO) format seemed the most promising possibility from the literature. In this format, three collocational pairs are presented to the test taker. One of these, however, is a non-collocation. Subjects are asked to select the two natural and frequently occurring collocations (COL) (Figure 7). One test item for each of the 36 target words was created in this format. We used only collocations in which the target words and their collocates retained their literal meanings (accurate description = a description that is accurate). The two acceptable collocation options also had to have a minimum Mutual Information (MI) score of 3, a commonly accepted threshold value in corpus linguistics to indicate strength of association (Schmitt, 2010), and a minimum frequency count of 10 in the COCA corpus (Davies, 2008). Selection of the two correct options was then guided by part of speech, because the options all needed to be from the same word class. Care was also taken to select component parts that were of a higher frequency than the targeted frequency band of the target word. This means that sometimes the top collocations had to be discarded for that reason. For instance, the top collocation for the word "behaviour" according to the outlined criteria is "aggressive." Using this pair, however, means that the second pair also had to consist of an adjective+noun collocation. This means that the words "animal" and "child", although among the top collocates of "behaviour" could not be used for the second pair. Neither did we want to use the word "abusive" because it is less frequent (3 K) than the word "antisocial" (2 K, according to the BNC-COCA list), even though it is an equally strong collocate of "behaviour." The non-collocations, while semantically plausible, were checked against the COCA data to make sure they did not occur as a partnership, or not more than once. The component parts of the non-collocations were also controlled for their individual frequency level.

To increase the number of participants and the generalizability of findings, we explored whether we could use a less time-intensive method of measuring meaning recall knowledge, compared to the interviews. Written meaning recall measures have already been successfully used in other studies on vocabulary item formats (e.g., Zhang, 2013), so we considered this option. We pretested a written measure

□ aggressive behaviour
□ antisocial behaviour
□ ugly behaviour

Figure 7. Collocation knowledge format (COL) (answers = *aggressive behaviour* and *antisocial behaviour*).

Write the meaning of each word in the space next to it. Describe it as precisely and in as much detail as possible. If you do not know the word at all, do not guess.

Example:

regret: I **regret** my decision.

feel sorry or sad about a mistake that I have made, and a wish that it could have been different and better

Figure 8. Written meaning recall measure (MR).

(the instructions of which and the example item shown to the students can be seen in Figure 8), where 21 Austrian EFL learners, who were similar to the eventual main study participants, were asked to describe the meaning of 40 randomly selected 3 K words and then were interviewed about the same words in an interview, as in the pilot study. For this we predefined the minimally required information that would have to be provided about the word's meaning by a participant and scored the answers from both measures accordingly. For instance, for the target word "bench," the concept of an object for seating and that this object seats more than one person because it is slightly longer or bigger than a chair were defined as the minimum required for a candidate to be awarded the full point. Candidates who wrote down the idea of an object for sitting down but did not mention anything about the object's size, were not credited with the full point. When probed in the interview, however, most candidates could describe the size of the object, which explains why the scores in the two measures were not completely identical. After outlying target words and participants were removed (10%), the correspondence between written measure and interview reached 88%. Although interviews clearly allow interactive probing of learner knowledge, it was felt that the more time-efficient written meaning recall measure (MR) provided essentially the same information and had the advantage of allowing the test to be administered to substantially greater numbers of participants.

Procedure

All measures were incorporated into a Web-based survey tool and administered online, creating four test versions. The procedure was piloted with 17 Austrian EFL learners from the target population, who did not participate in the main data gathering. After this, minimal changes were made to the instructions and one target word was replaced because it appeared too challenging for the population, even in the recognition formats. Students were presented with the test in the order shown in Table 4. This order was chosen to minimize cross-contamination between the tests because, for instance, administering the form-meaning tests before the derivative knowledge test would likely have resulted in the first test giving away many answers to the second test.

All participants started the test battery by answering the 36 items of the test of receptive derivative knowledge (1). Then, in a randomized fashion, each participant took one of the four form-meaning link knowledge tests (2), each of these containing all target items but in different formats as outlined in the pilot study section above. All participants then completed the same written meaning recall measure (3). After this, all participants answered the same 36-item collocation test (4). The participants were administered all measures in a one-hour session and took about 38 minutes on average to complete the test battery. Answers were afterwards coded 0 for incorrect answers and 1 for correct answers, with no partial credit given.

Results

The seven measures (Derivations, Collocations, Written Meaning Recall, and the four form-meaning link test versions) performed well in their reliability. The Cronbach alpha values are illustrated in Table 5.

Table 4. Order of tests in main study.

	Version A	Version B	Version C	Version D
1	Test of receptive derivative knowledge (DER)			
2	Form-meaning A	Form-meaning B	Form-meaning C	Form-meaning D
3	Written meaning recall measure (MR)			
4	Collocation test (COL)			

Table 5. Reliability indices of instruments in main study.

	DER	MR	COL	Form-Meaning A	Form-Meaning B	Form-Meaning C	Form-Meaning D
Cronbach's alpha	.84	.94	.96	.91	.93	.92	.92

The collocation test featured the highest Cronbach alpha value at .96; the alpha of the meaning recall measure was .94. The derivational knowledge measure yielded a lower, but still satisfactory, alpha of .84. The individual test versions' reliabilities, featuring nine items in each format (36 items in total), were between .91 and .93. Reliability within the formats ranged from an average alpha of .65 (MM) to .68 (MC), to .80 (CON), to .84 (DEF). This can be considered acceptable, given the lower number of participants per format (N = 24/25) compared to the derivative, collocation, and meaning recall measures (N = 99).

The facility values of the instruments also confirmed that the measures were appropriate in difficulty level because they seemed to provide a spread of scores. The DER subtest showed an average facility value of .64. The meaning recall test was slightly more difficult (.58) but still easier than the collocations test (.41). It is not surprising that the recognition formats were found to be easier for the candidates with average facility values of .77 for both the MM and the MC format, with the form recall formats being relatively more difficult (CON: .52; DEF: .45).

To explore the item format that best represented the verified meaning recall knowledge necessary for reading employability, student responses on the different formats were compared with their scores on the concurrent criterion measure of the written meaning recall. The results of matching and non-matching cases (N = 891, 99 candidates x 9 items per format) can be found in Table 6.

None of the item formats functions very well in estimating the meaning recall level of vocabulary knowledge necessary for reading. All formats show a mismatch between test item score and criterion measure score in around 23–26% of the cases. The MC format slightly outperforms the other formats with a matching rate of 77.3%, but this cannot be considered good enough to be very informative about reading employability.

While the overall performance of the recall and recognition formats was very different in the pilot study, the discrepancy in percentages of matching and non-matching cases between the formats is negligible in the main study's results. The item types all seem to perform fairly similarly on a general level (i.e., in matching the criterion measure). However, when considering the results in overestimation and underestimation, one can see that the formats behave very differently (Table 7).

As found in the pilot study, the recognition formats generally overestimate learners' meaning recall knowledge, whereas the form recall formats tend to underestimate it. In the MM format, overestimation of word knowledge at meaning recall level occurred in 22.2% of the cases, whereas the MC format overestimated the candidates' meaning recall knowledge in 20.3% of the cases. The definition form recall format underestimated candidates' meaning recall knowledge in about the same percentages of cases (19.1%). The form recall format with context performs somewhat unpredictably with almost as many cases overestimating meaning recall word knowledge (10%) as underestimating it (15.7%). Overall, it seems none of the formats investigated provide very trustworthy indications of the meaning recall level of lexical knowledge necessary for reading.

Table 6. Correspondence between item formats and the written meaning recall measure (overall percentage of cases).

	Form Recognition Multiple Matching	Meaning Recognition Multiple-Choice	Form Recall Definition-Only	Form Recall Sentence Context
Match	74.5%	77.3%	74.4%	74.3%
No match	25.5%	22.7%	25.6%	25.7%

Table 7. Correspondence between item formats and the written meaning recall measure (percentage of cases).

	Form Recognition Multiple Matching	Meaning Recognition Multiple-Choice	Form Recall Definition-Only	Form Recall Sentence Context
Match (point)	54.9%	56.7%	38.5%	42.2%
Overestimation	22.2%	20.3%	6.5%	10.0%
Underestimation	3.3%	2.4%	19.1%	15.7%
Match (no point)	19.6%	20.7%	35.9%	32.1%

We were also interested in whether the item formats could give information about learners' derivation knowledge, as indicated by the derivation test (Table 8). The table shows that no item format succeeds in adequately representing the derivational knowledge of the candidates. The various formats yielded very similar overall results, with only about two-thirds accuracy of the 891 cases (CON: 67.3%, MC: 64%, MM: 63.6%, DEF: 62.7%). One might argue that these form-meaning link knowledge items do not necessarily target or intend to measure derivational knowledge, but many studies have interpreted test results using these formats as indicating words are "learned" or "known," which presumably includes derivational knowledge about the target words. Our results suggest that this kind of interpretation is not tenable. This also calls into question the use of word family as a counting unit in vocabulary size tests (Kremmel, forthcoming). It seems that just because a learner can answer a form-meaning test item, this does not mean they know all of the members of that target word's word family, so interpreting a score as indicating full knowledge of a word family is questionable.

It also emerges that the form-meaning formats behave in a very individual fashion regarding their relationship with this type of word knowledge. In around 25% of the words tested through recognition formats, the candidates managed to answer the test item correctly without showing that they also knew the word's derivative forms in the derivation test. The opposite was the case in 28% and 22%, respectively, of the words tested in the recall formats (Table 9).

Furthermore, a comparison of the derivational measure scores with the meaning recall measure scores (3,564 cases = 99 candidates x 36 items) found that no clear inference about a person's knowledge of derivative forms can be drawn from their meaning recall knowledge either (Table 10). In 64% of the cases, candidates knew either both meaning recall and the derivational forms (43%) or neither (21%). However, there is a mismatch in 37% of the cases, which is in line with the mismatch percentages in Table 8. In 21% of the cases candidates could form the base word of the derivational variations without demonstrating knowledge of the meaning of that base word. In 16% of the cases, candidates knew the meaning of a word but could not connect the derivational forms to that base word in the derivation test. In sum, it seems that there are insufficient grounds to make substantial inferences about the derivational knowledge of a candidate from any type of form-meaning format.

We also compared the various item formats with the collocation measure. The representation of this knowledge aspect through the form-meaning item formats is even weaker than was the case with derivation, with scores matching in a maximum of 61.4% of the cases (DEF), and in only 59.1%

Table 8. Percentage of matching/mismatching cases comparing item formats and derivation test scores.

	Form Recognition Multiple Matching	Meaning Recognition Multiple-Choice	Form Recall Definition-Only	Form Recall Sentence Context
Match	63.6%	64.0%	62.7%	67.3%
No match	36.4%	36.0%	37.3%	32.7%

Table 9. Percentage of matching/mismatching cases comparing item formats and derivation test (detailed analysis).

	Form Recognition Multiple Matching	Meaning Recognition Multiple-Choice	Form Recall Definition-Only	Form Recall Sentence Context
Match (point)	52.2%	52.6%	35.5%	41.8%
Item>derivation	24.9%	24.4%	9.5%	10.4%
Item<derivation	11.4%	11.7%	27.7%	22.2%
Match (no point)	11.4%	11.3%	27.3%	25.6%

Table 10. Percentage of matching/mismatching cases comparing meaning recall knowledge and derivational knowledge.

		Meaning Recall Knowledge	
		Known	Not Known
Derivational knowledge	Known	43%	21%
	Not known	16%	21%

Table 11. Percentage of matching/mismatching cases comparing item formats and collocation test scores.

	Form Recognition Multiple Matching	Meaning Recognition Multiple-Choice	Form Recall Definition-Only	Form Recall Sentence Context
Match	51.7%	53.5%	61.4%	59.1%
No match	48.3%	46.5%	38.6%	40.9%

Table 12. Percentage of matching/mismatching cases comparing item formats and collocation test scores (detailed analysis).

	Form Recognition Multiple Matching	Meaning Recognition Multiple-Choice	Form Recall Definition-Only	Form Recall Sentence Context
Match (point)	35.2%	36.3%	22.8%	25.7%
Item>collocation	41.9%	40.7%	21.7%	26.5%
Item<collocation	6.4%	5.7%	16.8%	14.4%
Match (no point)	16.5%	17.3%	38.6%	33.4%

(CON), 53.5% (MC) and 51.7% (MM) for the other formats (Table 11). Again, a closer look at the matching and mismatching cases reveals differing format behaviour (Table 12). The participants very often were able to answer the recognition formats but not the collocation measure (41.9% and 40.7%). This was also the case with the form recall formats (21.7% and 26.5%), although they also produced many cases where the opposite occurred (16.8% and 14.4%), leading to considerable unpredictability about how the form-recall formats relate to collocation knowledge.

Discussion

To interpret a vocabulary size test score, one must know what the format shows about the examinee's ability to employ the target words on the test, or what the examinee knows about the target words. Most studies simply report that test takers "know" or have "learned" the vocabulary words answered correctly on a vocabulary test. Our results suggest that this type of reporting is far too simplistic. If a word is known or learned, this implies "employability," that is, the ability to employ that word in real communicative events in one or more of the four skills, for example, when reading and comprehending an authentic newspaper article. Certainly, teachers would wish to interpret vocabulary scores in this manner. Ideally, teachers interested in meaning recall knowledge should, of course, employ meaning recall formats. However, the unfeasibility of such formats and their introduction of additional problems, such as subjectivity and potential unreliability in scoring judgments, often makes teachers turn to the type of format investigated here. In our studies we explored the degree to which four common form-meaning formats could inform about the level of lexical mastery necessary to support employability in reading. Unfortunately, none of the formats investigated proved robust enough to be interpretable as representing meaning recall knowledge. Echoing the pilot findings, the main study showed that all four formats incorrectly indicated meaning recall knowledge in roughly 25% of the cases. Therefore, it seems that interpreting test scores (that use these formats) as showing learners can employ the target vocabulary in reading may be suspect.

Perhaps it is not surprising that we found that each item format had its own characteristics in this regard. If one wished to use the recognition formats (multiple matching and multiple-choice) to indicate meaning recall knowledge, there would be considerable overestimation, at around 20%. While this is not desirable, at least the mismatches were relatively systematic in one direction, because there was underestimation in only around 3% of the cases (an 89/11 split). This raises the possibility that some adjustment might be developed to reconcile the recognition and meaning recall scores. There has been considerable debate about correction formulas for yes/no tests (see, e.g., Huibregtse, Admiraal, & Meara, 2002; Stubbe, 2012; Stubbe & Stewart, 2012), and it seems possible that formulas might be developed for

MM and MC formats that would allow them to be satisfactorily interpreted as representing meaning recall (employability in reading) knowledge and perhaps employability in other skills as well.

Conversely, if one wished to use the form recall formats (DEF and CON) to indicate meaning recall knowledge, the interpretation would be more difficult. In about 26% of the cases, the formats did not accurately represent meaning recall knowledge. But the mismatches were not systematic. For the DEF format, there was underestimation in 19.1% and overestimation in 6.5% of the cases (a 75/25 split), but for the CON format, the gap was narrower (underestimation: 15.7%, overestimation: 10.0%—a 61/39 split). This makes it difficult to know whether the form recall format is underestimating or overestimating meaning recall knowledge, which would make the development of an adjustment formula extremely problematic. Although Laufer and Goldstein (2004) and Laufer *et al.* (2004) found that meaning recall items were easier than form recall items in general, it may be that the relationship between the formats for any particular target word is simply too variable to be used as a basis for consistent score interpretation.

There may also be other factors that affect how item formats can be interpreted. A comparison of results of the pilot study with findings of the main study indicates that the very low frequency of the target words in the pilot was problematic and partly caused the divergence between the recall and the recognition formats. Harking back to Laufer *et al.*'s (2004) findings, it seems that the gap in the strength of the form-meaning link knowledge opens up, particularly at the lower end of the frequency continuum. For example, the low-frequency word *skylark* (at the 15 K frequency level in the BNC-COCA), was known by 100% and 63%, respectively, in the recognition formats, but no candidate was able to recall the form of the word in either of the form recall formats. Schmitt (2014) found a similar phenomenon in his overview of the relationships between vocabulary size and a variety of depth measures. So score interpretation may depend not only on the item formats but factors like the frequency of the target words as well.

We have seen that interpreting scores from the four form-meaning item formats as the ability to *employ* the target vocabulary in language use may be questionable. But might the formats be more informative in examinees' *knowledge* of the target words? We explored this question by administering two concurrent tests of word knowledge measuring derivative and collocation knowledge. Unfortunately, the form-meaning formats did not indicate derivative or collocation knowledge to any great degree. This suggests that interpreting test scores based on these formats as showing that target words are "known" to a greater depth (i.e., having mastery of the word's derivative forms and collocations) is risky, with our results suggesting that the inference would be wrong in about 35–45% of the cases.

These results have at least four implications. First, they suggest that users of vocabulary size tests must be more careful about how they interpret and report the resulting test scores. Rather than assuming words are known or learned in a general way if items are answered correctly, users must consider what inferences are actually warranted about the learners' underlying lexical knowledge based on particular item formats.

A second related implication is that vocabulary test developers need to explicitly state what score interpretations are warranted for their tests, based on the formats they use. That is, they should "explicitly state what correct answers on their tests entail and what degree of depth they represent" (Schmitt, 2014, p. 943). At the moment, test users are typically left to interpret test scores as they wish, and it is almost inevitable that different users will interpret the scores in different ways. The only way the field can come to a better understanding of what various item formats tell us about lexical employability and knowledge is for test developers to include a prominent "score interpretation" element into their future validation procedures.

Third, if one wishes to make valid score interpretations for a number of knowledge aspects, it may be necessary to test those aspects separately. It might turn out that any particular item format is intrinsically limited in the amount of employability or knowledge it can realistically represent. While this notion is not novel or surprising, it should once again encourage caution about interpreting vocabulary size test scores as simply how many words a learner knows, without clearly outlining what "knowing" means.

Fuller descriptions of vocabulary knowledge may well require a battery of vocabulary formats, which can prove unwieldy (Read, 2000), although computer adaptive testing techniques may go some way toward making this approach more practical in the future.

Fourth, the results of the comparison between derivation test scores and meaning recall knowledge seem to suggest that knowledge of a word family member's meaning cannot be taken to also mean receptive knowledge of other word family members. Participants in this study could not consistently make the connection between derivational forms of a word and its base form, even though they demonstrated knowledge of the meaning of that base word. This echoes findings of earlier studies (Schmitt & Zimmerman, 2002; Ward & Chuenjundaeng, 2009) and further questions the use of word family lists as a sampling basis for vocabulary tests. At a minimum this suggests that the vocabulary size estimates from established tests such as the VLT or the VST (which use the word family counting unit) are probably overly optimistic. The evidence presented in this article could even be seen as an argument for abandoning word family lists altogether as the sampling basis for vocabulary size tests. It is unclear what would replace word families, although lemma lists are the obvious counting unit to begin researching in the first instance (Kremmel, forthcoming).

Conclusion

Clearly, this is only initial research, given the limited numbers of target words, item formats, and participants, who were mainly from one L1 population. However, the results do provide evidence for the need to be more careful when interpreting scores from vocabulary size tests. As Schmitt (2014, p. 943) points out, "Nobody interprets the [vocabulary test] scores as simply words that learners can answer on a vocabulary test." The present findings suggest that the four form-meaning item formats should neither be readily interpreted as providing accurate information about L2 learners' ability to employ the target vocabulary in reading (RQ1), nor about their knowledge of derivative forms (RQ2) or collocations (RQ3). It remains to be discovered exactly what score interpretations are justified for any particular vocabulary item format.

References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency*. London, UK: Continuum.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, UK: OUP.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, UK: OUP.

Barfield, A., & Gyllstad, H. (2009). *Researching collocations in another language: Multiple interpretations*. Basingstoke, UK: Palgrave Macmillan.

Davies, M. (2008). *The Corpus of Contemporary American English: 450 million words, 1990-present*. Retrieved from http://corpus.byu.edu/coca/

Eyckmans, J. (2009). Towards an assessment of learners' receptive and productive syntagmatic knowledge. In A. Barfield, & H. Gyllstad (Eds.), *Researching Collocations in another language: Multiple interpretations* (pp. 139–152). New York, NY: Palgrave Macmillan.

Goodrich, H. C. (1977). Distractor efficiency in foreign language testing. *TESOL Quarterly*, *11*(1), 69–78. doi:10.2307/3585593

Grabe, W. (2009). *Reading in a second language*. Cambridge, UK: Cambridge University Press.

Henning, G. (1991). A study of the effects of contextualization and familiarization on responses to the TOEFL vocabulary items. *TOEFL Research Reports, 35*. Princeton, NJ: Educational Testing Service.

Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes–no vocabulary test: Correction for guessing and response style. *Language Testing*, *19*(3), 227–245. doi:10.1191/0265532202lt229oa

Kamimoto, T. (2008, September 11–13). *Guessing and vocabulary tests: Looking at the Vocabulary Levels Test*. Paper presented at the 41 Annual BAAL Conference, Swansea, UK.

Kremmel, B. (forthcoming). Word families and frequency bands in vocabulary tests: Challenging conventions. *TESOL Quarterly*.

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, *21*(2), 202–226. doi:10.1191/0265532204lt277oa

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

Martinez, R. (2011, April). *Putting a test of multiword expressions to a test*. Presented at the IATEFL TEASIG conference, University of Innsbruck, Innsbruck, Austria.

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

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Florence, UK: Heinle and Heinle.

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

Nation, I. S. P. (2012a). *Vocabulary size test instructions and description*. Retrieved January 7, 2015 from https://www.victoria.ac.nz/lals/about/staff/paul-nation. Last revised on October 23, 2012.

Nation, I. S. P. (2012b). *The BNC/COCA word family lists*. Retrieved January 7, 2015, from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Information-on-the-BNC_COCA-word-family-lists.pdf

Nation, I. S. P. (2014). *Vocabulary Size Test (monolingual 20,000, Versions A & B)*. Retrieved February 3, 2014, from https://www.victoria.ac.nz/lals/about/staff/paul-nation

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13.

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

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam, The Netherlands: John Benjamins.

Paul, P. V., Stallman, A., & O'Rourke, J. P. (1990). *Using three test formats to assess good and poor reader's word knowledge*. Technical Report No. 509 of the Center for the Study of Reading. Champaign, IL: Center for the Study of Reading, University of Illinois at Urbana-Champaign.

Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring Yes–No vocabulary tests: Reaction time vs. nonword approaches. *Language Testing*, *29*(4), 489–509. doi:10.1177/0265532212438053

Pike, L. W. (1979). *An evaluation of alternative item formats for testing English as a foreign language* (TOEFL Research Report 2). Princeton, NJ: Educational Testing Service.

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

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

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, *64*(4), 913–951. doi:10.1111/lang.2014.64.issue-4

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.

Schmitt, N., & Zimmerman, C. B. (2002). Derivative word forms: What do learners know? *TESOL Quarterly*, *36*(2), 145. doi:10.2307/3588328

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

Stewart, J., & White, D. (2011). Estimating guessing effects on the vocabulary levels test for differing degrees of word knowledge. *TESOL Quarterly*, *45*(2), 370–380. doi:10.5054/tq.2011.254523

Stubbe, R. (2012). Do pesudoword false alarm rates and overestimation rates in yes-no vocabulary tests change with Japanese university students' ability levels? *Language Testing*, *29*(4), 471–488. doi:10.1177/0265532211433033

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

Ward, J., & Chuenjundaeng, J. (2009). Suffix knowledge: Acquisition and applications. *System*, *37*(3), 461–469. doi:10.1016/j.system.2009.01.004

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, *30*, 79–95. doi:10.1017/S0272263108080042

Zhang, X. (2013). The I don't know option in the vocabulary size test. *TESOL Quarterly*, *47*(4), 790–811. doi:10.1002/tesq.98