*Routledge Handbook of Vocabulary Studies*
Revised version – Feb 2018

# 32. Key issues in measuring vocabulary knowledge

John Read
The University of Auckland

## INTRODUCTION

This chapter focuses on some ongoing issues in designing and administering instruments to measure the vocabulary knowledge of second language learners. As such it is intended to complement the other chapters in this volume which deal with more specific aspects of vocabulary assessment. The work discussed here is concerned primarily with knowledge of individual words rather than, say, competence with multi-word lexical units or the ability to use vocabulary for communicative purposes.

There has been a strong tendency in L2 vocabulary studies for certain tests to become recognised as standard measures for a variety of uses. For many years Nation's Vocabulary Levels Test (VLT), first in the original form (Nation 1983) and then in its revised version (Schmitt, Schmitt & Clapham 2001), was widely seen as the best way to assess vocabulary size, or breadth of vocabulary knowledge – even though it was not designed to give an overall estimate of the total number of words known. This limitation was rectified to some extent by the development of the Vocabulary Size Test (VST) (Nation & Beglar 2007), which sampled systematically across a range of word frequency levels and was shown to have very good measurement qualities, at least when it was administered to Japanese learners with different degrees of proficiency in English (Beglar 2010). To a more limited extent, a series of tests developed by Meara and his associates based on the Yes/No test format (Meara & Jones 1988; Meara & Miralpeix 2016) have performed a similar, though less prominent role.

The same situation applies to measures of depth of vocabulary knowledge. As Schmitt (2014) notes, studies of depth have most often employed tests based on Read's (1993, 1998) word associates format, which assesses the ability to match words related in various ways that are assumed to reflect the organisation of the mental lexicon. Another popular instrument is the Vocabulary Knowledge Scale (VKS) (Wesche & Paribakht 1996), which requires learners to rate their knowledge of a set of target words and provide some confirming evidence if they do claim to know a word.

However, the attention given to this small number of high-profile instruments can obscure two points: one is that there are various purposes for setting out to assess learners' vocabulary knowledge; and the other is that there is a wider range of test formats available that just these few. Thus, first it is useful to clarify is what the purposes for measuring vocabulary knowledge are: why do we want or need to use vocabulary tests and other forms of assessment?

**Purposes of Assessment**

One goal, broadly stated, is to measure vocabulary size, which is also referred to as breadth of vocabulary knowledge. Tests like the VLT and VST have been designed to sample words across a wide range of frequency levels, and the number of correct answers to the test items is used to estimate the learner's vocabulary size. A true size test like the VST sets out to obtain a measure of all the words that someone knows, which is conceptually challenging if not impossible – except perhaps in situations where a learner's exposure to the second language is restricted to the classroom and the textbook. The sampling frame for size tests will always be limited by the scope and representativeness of the corpora from which the word frequency data are derived. In addition to that, learners with relatively limited vocabulary knowledge may get frustrated and engage in blind guessing if the test contains a lot of unknown lower

frequency words which they are required to work through just for the sake of finding the odd one that they do know. The alternative is a test like the VLT, which focuses on the range of frequency levels where the words which learners know – or *need* to know – are most likely to be found. Thus, Nation (2016) makes a generic distinction between a size test and a levels test, the latter having the more modest goal of establishing how many of the three, or five, thousand most frequent words are known, rather than estimating total size.

But, to come back to purpose, why do we want to measure vocabulary size, in either an absolute or relative sense? Some distinct assessment purposes can be identified:

- To obtain a broad measure of proficiency in the language for diagnostic purposes, as in DIALANG, the web-based diagnostic assessment system (Alderson, 2005).
- To place students in classes in a language school (Harrington & Carey, 2009).
- To evaluate the adequacy of learners' vocabulary knowledge for specific communicative purposes (Schmitt, Jiang & Grabe 2011).

However, it is often not clear which of these purposes is being addressed in research studies on vocabulary breadth tests – a point I will return to below in the section on validation.

Given the prominence of tests like the VLT and the VST, the goal of measuring vocabulary size tends to dominate the L2 vocabulary literature, but it is important to identify other purposes of measuring vocabulary knowledge, which may not necessarily be mutually exclusive.

- To measure the outcomes of a period of vocabulary learning. This may be quite brief, as in the results of an experiment to investigate, say, the effects of presenting a set of words in particular contexts or groupings within a single study session in a controlled environment; or it may be an extended period of study in a more naturalistic

classroom setting. In this case, the testing will normally focus on the specific vocabulary items that the students were supposed to study. Particularly in the case of a controlled experiment, there may be pretesting as well as posttesting (see Kremmel, this volume)

- To measure specific language learning skills. So far there have been few published reports of such instruments, apart from ones used in specific studies, such as Schmitt and Zimmerman's (2002) Test of English Derivatives (TED). This research gap has been recently addressed more systematically by Sasao (Sasao & Webb 2017, 2018) with the development of his Word Part Levels Test and Guessing from Context Test (see Sasao, this volume).
- To gain a better understanding of the mental lexicon and of vocabulary learning processes. Paul Meara and his students and colleagues at Swansea have developed multiple tools for this purpose (Meara & Miralpeix 2016) to investigate quite an array of research questions (Fitzpatrick & Barfield 2009).

There are other purposes, such as to assess the ability of learners to use vocabulary productively in speech and writing, and to evaluate the contribution of vocabulary knowledge to the performance of communicative tasks, which go beyond the scope of this chapter (see Kyle, this volume). For further discussion of assessment purposes, see the section below on validation, and also Nation (2013, Chap 13).

## CRITICAL ISSUES AND TOPICS

### Choice of Test Formats

In terms of test design there appear to be a limited number of item formats available to measure vocabulary knowledge. As in other areas of educational assessment, the predominant mode of delivery has been a group-administered, paper-based test in which the stimulus material is presented – and the learners respond – in written form. The advent of computer-based testing has opened up new opportunities: for learners to take tests individually, and at remote locations; for tests to be individualized to a learner's ability level, using computer-adaptive procedures; for reaction times and other elements of response behaviour to be recorded; and for the results to be scored and analysed automatically. However, the basic item types tend to remain the same. Although the sophisticated technology available to the major test publishers has the potential to offer a wider range of assessment formats, most smaller scale computerized tests rely on low-tech authoring software such as that available through learning management systems like Blackboard, Canvas and Moodle, with their restricted number of item types, which are largely the same as those used in most paper-based tests. These item types lend themselves well to conventional means of measuring vocabulary knowledge.

Let us review those standard item types. First, it is useful to distinguish between selected-response items, where the learners select a response from two or more options given, and constructed-response ones, where the learners write their own response in the form of a word, a phrase, a sentence and so on.

This distinction fits neatly with the traditional division in vocabulary studies between receptive and productive vocabulary knowledge, although I much prefer the terms recognition and recall, for reasons I have given elsewhere (Read 2000: 155-156). Thus, selected-response items require the learners to show that they can recognise the correct

meaning or form of a given target word, whereas constructed-response items set the more demanding task of being able to recall the meaning or form.

Selected-response items (recognition knowledge)

The multiple-choice format continues to be the classic way to measure recognition knowledge. It has gained new prominence through its use in the original Vocabulary Size Test (Nation 2012; Nation & Beglar 2007; Beglar 2010) and the various versions derived from it. In this case, the format is employed to assess meaning recognition: the test-takers select which of four options best represents the meaning of the target word presented in a short, non-defining sentence in the stem of the item:

circle: Make a <circle>.
a rough picture
b space with nothing in it
c round shape
d large hole

Multiple-choice items can also be used to measure form recognition, as in this example from an experiment by Webb (2007) to test knowledge of how to spell the pseudoword *denent*:

1. (a) denant (b) danant (c) danent (d) denent

A second widely used format for meaning recognition is a matching task. The canonical example of an instrument of this kind is the Vocabulary Levels Test (Schmitt, Schmitt & Clapham, 2001), where learners are presented with sets of three definitions and five words, and need to match each definition with the word it refers to. The same format has been retained in the new computer-based Vocabulary Levels Test (Webb, Sasao & Balance 2017).

In these tests the words in each set belong to the same part of speech but are quite different in meaning, so that the test provides what Nation (2013: 536-7) calls a 'sensitive' measure of vocabulary knowledge, in that it does not require the test-takers to distinguish between semantically related words. The format could also be used in less sensitive tests to assess the ability of more advanced learners to make such distinctions.

Of course, in a sense all selected-response test items involve matching of some kind (including the multiple-choice format), but the label is conventionally applied to the type of test just described, where there are clusters of target words and definitions. Another variation on the matching theme is represented by the word associates format (Read 1993) where the test-takers select words that are semantically related to a given target word, as in this example:

*team*

| | | | |
|---|---|---|---|
| alternative | chalk | ear | group |
| orbit | scientists | sport | together |

The intended responses in this case are *group*, *scientists*, *sport* and *together*. This again can be seen as a sensitive test, where the distractor words are not related to the target item. A less sensitive version of the format is this adaptation for primary school students in the Netherlands (Schoonen & Verhallen 2008):

| | | |
|---|---|---|
| | fruit | monkey |
| nice | **banana** | (to) slip |
| | peel | yellow |

Although all six associates have a connection to the target word, *fruit*, *peel*, and *yellow* are judged to constitute a more sophisticated definition of what it means.

The matching principle has also been applied to tests of collocational knowledge, assessing the ability of learners to recognise the correct combination of words, as well as what the multi-word unit means, as in this sample from Revier's (2009) CONTRIX format:

| | | | |
|---|---|---|---|
| The quickest way to win a friend's trust is to show that you are able to _____ | tell | a/an | joke |
| | take | the | secret |
| | keep | ---- | truth |

In short, then, items based on matching require learners to demonstrate their vocabulary knowledge by identifying connections between words and their meanings, or between words and other words – all of which are provided as an integral part of the test format.

A third type of selected-response item is found in the Yes/No test, which presents the test-takers with a series of words and simply asks them to report whether they know the meaning of each one or not. Thus, it can be regarded as a kind of self-assessment task, which provides only indirect evidence of the validity of the test-takers' responses by including a proportion of non-words (also called pseudowords) so that individuals' scores can be adjusted downwards if they respond 'Yes' to one or more of the non-words.

A big issue, though, is the opportunity for test-takers to guess answers from the options provided, as discussed further below. This forms part of the rationale for the alternative approach of requiring to supply their own answers to the test items.

Constructed-response items (recall knowledge)

The essence of a constructed-response test item is that the test-takers are prompted to recall the meaning or form of a target vocabulary item. In the simplest case, they are given a list of L2 words and must supply an expression of the meaning of each one: an L2 definition or synonym or an L1 equivalent. With some framing the same kind of item can be employed to elicit various other aspects of word knowledge: a word's part of speech (*undertake*: verb), a

derived form (*receive* → *reception*), or a collocate (*to make* + *a decision*). A more demanding task is for the learners to produce a whole sentence containing each target word. Since it is also difficult to mark, this type of task is not normally included in formal, large-scale tests.

Thus, a more common type of constructed-response item involves completing a gap in a sentence. The idea is that the sentence is written to create a context which allows the test-takers to infer what the missing word is (and where appropriate, what its grammatical form should be), as in these examples:

He worked in this office for a _____ of four years. [period]

Many houses were completely _____ in the earthquake. [destroyed]

It is often the case that more than one word can fill the gap. In a classroom progress test, the teacher may choose to accept only a word which the learners have recently studied. Another way to signal the *intended* answer is to supply the first letter of the word. One well-known test in this format, Laufer and Nation's (1995) Productive Levels Test, goes further than that. Since the test was intended to be the 'productive' counterpart to the original Vocabulary Levels Test (Nation 1983), the authors chose to supply up to six or seven initial letters in order to ensure that only the target words in the VLT were elicited, so that such items in effect involve more recognition that recall. This highlights the point that constructed-response items are inherently less controlled, and may be more difficult to score, than selected-response items.

One test which combines the two types of item is the Computer Adaptive Test of Size and Strength (CATSS) (Laufer & Goldstein 2004). It is based on two distinctions:

- supplying the form for a given concept vs. supplying the meaning for a given form; and

- recall vs. recognition (of form or meaning) (Laufer et al. 2006: 206).

This gives rise to four 'degrees of knowledge', each with a corresponding type of test item. Using Rasch analysis, Laufer et al. (2006) showed that the two types of recognition knowledge (of form and of meaning), which were both tested in the CATSS by means of multiple-choice items, were quite easy, whereas recall knowledge of meaning was significantly more challenging, and recall knowledge of form was most difficult of all. The results confirmed the general expectation that constructed-response items are more demanding for learners than selected-response items, and as a general rule vocabulary tests employ one of these types of item, rather than combining them in the way that the CATSS does.

**Role of L1 and Bilingual Testing**

Another issue is the role of the learners' first language in vocabulary testing. There is the obvious consideration of whether the intended population of test-takers share a first language, in which case their own language represents an obvious resource for communicating the meaning of L2 vocabulary, whether it be through recognition or recall. This applies, for instance, to learners of English in Japan, learners of Japanese in Brazil, or learners of Spanish in New Zealand – although the use of L1 may be constrained if the teacher is not fluent in the students' language. On the other hand, the influential English vocabulary tests in the international literature have grown out of educational contexts in English-speaking countries where learners from different language backgrounds study together in the same class or school and where the teacher may not speak any of their languages. In those situations a monolingual test in English is often seen as the only practicable option.

To discuss the appropriate roles of L1 in the second language classroom more broadly is beyond the scope of this chapter. It is a matter of ongoing debate within both the mainstream literature on language teaching methodology (see, eg, Du 2016; Turnbull & Dailey-O'Cain 2009), and the more recent work on translanguaging in bilingual and multilingual classrooms (Cenoz & Gorter 2015; Garcia & Li 2013). Within these diverse contexts appropriate forms of vocabulary assessment will find their place.

Research on L2 vocabulary testing has paid comparatively little attention to the L1 versus L2 question until recently. Some researchers in Japan (e.g. Stubbe 2015) have noted that one practical argument in favour of using the Yes/No format rather than a translation (L2 to L1) task for testing word knowledge is that it is time-consuming to mark manually the variety of L1 equivalents that the learners produce in the latter task. On the other hand, Schmitt (2010) argues that an individual face-to-face interview is '[p]erhaps the best way of determining "true" underlying knowledge' of vocabulary (p. 182) and, in research studies he has co-authored (Gyllstad, Vilkaite & Schmitt 2015; Schmitt, Ng & Garras 2011), at least some of the interviews have been conducted in the learners' L1.

Starting in the 1990s Paul Nation encouraged the development of bilingual versions of the Vocabulary Levels Test, with the definitions translated into various L1s, and made them available on his website. The argument was that the bilingual version of the test was a purer measure of L2 word knowledge because it reduced the reading load involved in comprehending the definitions (Nation 2001: 351). However, there was no published work to evaluate any of the bilingual versions. This has changed with the advent of the Vocabulary Size Test (VST). Not only are there bilingual versions in various Asian languages (Mandarin, Japanese, Korean, Thai, Gujarati, Tamil), but also several validation studies have been published, for the Vietnamese (Nguyen & Nation 2011), Persian (Karami 2012), Russian

(Elgort 2013) and Mandarin (Zhao & Ji 2016) versions. Both Nguyen & Nation (2011) and Karami (2012) found that their bilingual versions significantly distinguished between university student learners at three levels of proficiency, and there was a broad (though inconsistent) pattern of declining scores from the high frequency to lower frequency sections of the test. Zhao and Ji (2016) obtained similar results using Rasch analysis for a reduced 80-item version of the Mandarin test, covering the first 8000 frequency levels.

In her study with intermediate-level Russian learners, Elgort (2013) was able to compare the performance of the two versions of the test. Each learner took 70 bilingual items and 70 monolingual ones. Although the difference in mean scores was not great (32.97 vs. 29.61 respectively), the items with Russian options were significantly easier than the monolingual items (with a large effect size). Further analysis showed that learners with lower scores overall gained more benefit from the bilingual options, as compared to higher scoring learners.

**The Effects of Cognates**

A related issue, arising particularly from Elgort's work, is the role of cognates and loanwords in test performance. Researchers have long been aware of the effects of cognates when the L1 and L2 have been related European languages. In his study of the Vocabulary Levels Test Read (1988) noted that some Spanish-speaking students did not follow the general pattern of progressively lower scores at lower word frequency levels because of their knowledge of cognates at those levels. Similar results have come out of research with francophone students in Canada. Meara, Lightbown and Halter (1994) found that a group of these students scored significantly higher on a Yes/No test of English vocabulary in which half the target words were French-English cognates than one in which cognates were excluded. In a larger scale study of university placement test results, Cobb (2000) showed that the relatively high scores

gained by francophone students on the 2000 word level of the VLT masked a lack of knowledge of high-frequency words of Anglo-Saxon (or Germanic) origin. The students scored noticeably higher on items where the target word and its definition had cognates in French.

Returning to Elgort's (2013) research on the VST with Russian learners, one of her findings was that 34% of the target words in the test had Russian cognates, which was significantly higher than the estimated 27% of cognates in the language as a whole. Since responses to cognate target words were significantly more accurate, this suggested that the test may have overestimated the students' vocabulary size by as much as 1000 word families. Thus, one further advantage of a version of the test designed for students with a particular L1 is that the proportion of cognate words can in principle be controlled, whereas it is a confounding variable when the monolingual VST is administered to learners from different language backgrounds.

The influence of English as a global language in the modern world means that many non-European languages, which have no genetic relationship to English, have expanded their lexicon by borrowing words from English. As has often been observed, this is a particularly salient feature of modern Japanese vocabulary (Daulton 2008). Although linguistically cognates and loanwords have different origins, their potential for either facilitating L2 vocabulary acquisition or causing confusion through 'false friends' is similar, with the result that the two terms tend to be used interchangeably in the current L2 literature.

In an L2 to L1 translation task for Japanese learners, Jordan (2012) found that beyond very familiar vocabulary at the 1000-word frequency level non-cognates were significantly more difficult to translate than cognates. Similar results emerged from a study by Laufer and McLean (2016) involving students in Japan and Israel who took the new Computer Adaptive

Test of Size and Strength (CATSS). A version of the test which included loanwords in the respective L1s of the learners was easier than one which excluded such words, at least for the items in the test which required active or passive recall rather than just recognition of the words. In addition, the positive effect of loanwords on the test scores was more evident for lower proficiency learners than those with advanced vocabulary knowledge. In a recent study in Israel using the same test, Laufer and Levitsky-Aviad (2018) largely confirmed the results of the Laufer and McLean study. They found very little difference between a version with an uncontrolled number of loanwords and one with a proportionate number, but the actual numbers were small: 6 versus 3.

Thus, there is some evidence that the presence of cognates or loanwords as target items in a vocabulary test can affect the test-takers' performance, especially for elementary and intermediate level learners. This is not to say that such words should be omitted, but their possible influence should be taken into account, particularly in tests of vocabulary size administered to students with a single first language. If the proportion of loanwords is not controlled, it may lead to overestimates or underestimates of the number of words that the learners know within the word frequency range covered by the test.

**Guessing and Confidence**

With any kind of selected-response test format there is the potential for test-takers to increase their score by guessing, without knowing what the correct answer should be to some or more of the items. It is important to distinguish between blind (or 'wild') guessing and that based on some relevant knowledge. Traditionally, formulas to correct for guessing have been applied to various types of educational tests to discourage test-taker from guessing blindly;

however, measurement experts (eg, Ebel & Frisbie 1986: 215-218) have tended to question their validity and effectiveness.

In vocabulary assessment, Nation (2001; 2013) has consistently argued that students should be encouraged to draw on partial knowledge of vocabulary items, even if they are not sure of what the correct answer is. However, in his research on the word associates format, Read (1993, 1998) observed that some learners were unwilling to guess unless they were confident about an answer and thus were more conservative in their responding behaviour than others were. This has been a significant focus of the research on the Yes/No format, and a number of scoring formulas have been devised to adjust the final scores according to the number of false alarms (Yes responses to nonwords) (Huibregtse, Admiraal & Meara 2001; Mochida & Harrington 2006).

In the first instance, if guessing is an issue, it is desirable to include a statement in the instructions about whether it is acceptable to guess answers, and certainly there should be a warning if a penalty is to be imposed on incorrect responses. In a Yes/No test it is possible to add a 'not sure' response category, with such answers being treated as No for scoring purposes. The V_YesNo test uses Yes and Next as its two response buttons (Meara and Miralpeix 2016), presumably to encourage test-takers to move on to the next item if they are not sure whether they know the currently displayed word. However, regardless of what is stated in test instructions, it is also important to investigate further ways in which to communicate the nature of the task to the test-takers

Jeffrey Stewart and his colleagues in Japan have conducted a series of studies on the effects of guessing in selected vocabulary tests. Stewart and White (2011) used elementary probability theory and multiple simulations to calculate the likelihood that guessing inflates the scores on a selected-response test like the VLT. The results showed that, if test-takers are

assumed to know up to 60% of the target L2 vocabulary, their VLT scores will be boosted nearly 17% by guessing. At higher levels of vocabulary knowledge, the effect is diminished by ceiling effects. The authors argue that constructed-response items will yield more accurate estimates of vocabulary size. Subsequently, Stewart (2014) published a similar critique of the VST, making the case that the probability of guessing correctly without knowing the target word is even higher with four-option multiple-choice items. He points out that Rasch analysis – as used by Beglar (2010) in his influential validation study of the VST – cannot detect systematic patterns of guessing in a whole test-taker population, as distinct from the inconsistent patterns of individual test-takers. Using the three-parameter logistic model instead of Rasch to analyse a large set of VST scores, McLean, Kramer & Stewart (2015) found that a large proportion of the variance in scores for lower frequency words was attributable to guessing rather than actual knowledge of the words.

Research on the Vocabulary Size Test (VST) has investigated the effects of adding a fifth 'I don't know' (IDK) option to each of the multiple-choice items. Zhang (2013) compared the original VST with a modified version containing this option. A meaning recall task, which required the test-takers to give the meaning of each word in L1 or L2, was used to help identify whether they had at least some partial knowledge of the target word in responding to each VST item. The IDK option significantly reduced both types of guesses, especially when it was accompanied by an explicit warning of a penalty for wrong guesses. The penalty warning in particular had the effect of discouraging learners from guessing on the basis of partial knowledge. However, this study did not take into account the fact that test-takers select the IDK option to varying degrees. Using computer simulations, Stoeckel, Bennett and McLean (2016) demonstrated that variable use of the IDK option was an extraneous factor which could significantly affect the validity of VST scores.

Selected-response test formats will continue to have a major role in measuring vocabulary knowledge because of their practical advantages, especially in large-scale tests. This means that we need to keep investigating the effects of guessing on test performance among particular populations of learners and to evaluate the effectiveness of the various strategies to discourage blind guessing while at the same time encouraging learners to demonstrate partial knowledge of the target vocabulary items.

**Validation of Tests**

A recurring issue with tests of vocabulary knowledge is how to validate them. There is a long tradition of using some kind of criterion measure of vocabulary knowledge, especially with selective-response tests. As we have seen, in small-scale studies it is possible to conduct individual interviews with learners to elicit what they 'really' know about the target words and also to obtain retrospective verbal accounts of the strategies they employed in producing answers to the test items (see, eg, Gyllstad, Valkaite & Schmitt 2015; Read 1993; Schmitt, Ng & Garras 2011). On a larger scale, another popular criterion for validating tests based on selected-response formats is to use a written constructed-response task which requires the learners to provide their own synonym, definition or L1 equivalent for each target word (see eg, Mochida & Harrington 2006; Zhang 2013). Going beyond single criterion measures, the study of the VLT by Schmitt, Schmitt & Clapham (2001) represented a kind of high-water mark in terms of the range of evidence assembled to validate the test as a profile of the test-takers' vocabulary knowledge.

An alternative approach was adopted by Beglar (2010) for the validation of the original Vocabulary Size Test (see also McLean, Kramer & Beglar 2015). This approach draws primarily on Rasch analysis to generate evidence for the various components of an enhanced

version of Messick's influential framework for test validity. It is true that well-designed vocabulary tests typically have high reliability and other positive measurement qualities which are reflected in the Rasch statistics. However, the approach has been criticised as the basis for justifying estimates of vocabulary size derived from VST scores (Gyllstad et al. 2015; Stewart 2014) because it is not sensitive to systematic guessing behaviour among the test-takers.

Another weakness in the validation of vocabulary tests goes back to the introductory comments at the beginning of the chapter about the dominance of a small number of tests which have been promoted as one-size-fits-all instruments for use with learners worldwide. This is changing in the case of the VST, with the appearance of multiple versions of the test, including various bilingual versions, and debate over whether learners should take the whole test or just the items representing the higher frequency vocabulary they are likely to know. In addition, research on tests using the Yes/No format have shown rather different response behaviours by test-takers in different parts of the world. Thus, it is desirable to see test validation as an ongoing process of justifying the use of tests and test formats in particular linguistic and educational contexts.

Modern test validity theory also requires that tests should be validated for specific purposes. In numerous studies of vocabulary tests, it is not very clear what purpose they are designed to serve. Among the studies reviewed in this chapter, several instruments have been designed for placement purposes (Fountain & Nation 2000; Harrington & Carey 2009; Meara & Jones 1988). The original VLT (Nation 1983) was presented as a diagnostic tool for teachers in the classroom. In the research by Harrington and his colleagues with the Timed Yes No test (see below) in Australia and Oman, the purpose of the test is clearly stated and an appropriate external criterion measure (IELTS scores, grade point averages) was chosen. On the other

hand, in discussing the validation of the VST, Beglar (2010) claims a whole range of uses for the test, from monitoring learner progress in the classroom through the achievement of curriculum objectives to understanding 'the impact of educational reform on vocabulary growth' (p. 210). Although it could well be that a robust instrument like the VST can be used in all these different ways, its value for each of these purposes should be empirically evaluated.

## FUTURE DIRECTIONS

### Oral Vocabulary

The great preponderance of research on assessing vocabulary knowledge focuses on words in their written form, in keeping with the traditional relationship between vocabulary studies and research on reading in both L1 and L2. Perhaps it also reflects the fact that vocabulary study is popular in foreign language learning environments where there is limited exposure to the spoken language.

One early initiative to assess knowledge of spoken vocabulary was the graded dictation test developed by Fountain in the 1970s (Fountain & Nation 2000). The dictation text was carefully constructed to incorporate words from progressively lower frequency levels. Although test-takers write the whole text, only the target words are scored. A small-scale validation study showed a correlation of .78 with the {written) Vocabulary Levels Test, but there was no independent measure of the learners' ability to understand the individual words in their spoken form.

Other aural vocabulary tests have used the Yes/No format, presenting the target words in isolation. James Milton and his colleagues created AuralLex as an oral version of the written

X-Lex test to investigate the development of what they refer to as phonological and orthographic vocabulary size. Milton and Hopkins (2006) observed that, after the initial period of language learning, orthographic word knowledge tended to outstrip phonological knowledge, particularly among advanced learners. This pattern was more marked among Greek-speaking learners than Arabic speakers, whose vocabulary size in both modes of input was relatively low. From a follow-up study, Milton and Riordan (2006) concluded that growth in phonological vocabulary among Arabic speakers was constrained by inefficient reading strategies transferred from L1. More recently, Milton, Wade & Hopkins (2010) correlated scores on the two vocabulary tests with IELTS band scores. The correlations with the overall IELTS scores were quite substantial: .68 for X_Lex and .55 for AuralLex. As might have been expected, X-Lex had a stronger relationship with the IELTS reading and writing bands, whereas AuralLex correlated well with the listening and speaking scores.

McLean, Kramer and Beglar (2015) created a Listening Vocabulary Levels Test, using the multiple-choice item format of the VST but covering only the first 5000 words of English plus the Academic Word List. The stem of each item was spoken in English, with the response options written in Japanese to minimise the reading load. The scores on the 150-item test were moderately correlated (r=.54) with Parts 1 and 2 of the TOEIC listening test. In validating the test, the researchers employed the same Rasch-based framework adopted for Beglar's (2010) evaluation of the original VST, along with retrospective evidence from face-to-face interviews with a sample of the test-takers. This qualitative evidence showed that blind guessing appeared to play only a small role in the learners' response behaviour.

Thus, as with research on spoken vocabulary generally, there is a great deal of scope for further development of suitable tools for measuring knowledge of words in their oral form. The nature of speech makes it more challenging to hear and comprehend spoken words,

especially if they are presented out of context, as words commonly are in their written form. This leads to the further question of whether words that learners can identify in isolation can be comprehended in a discourse context as well. In her research, van Zeeland (2013) found a substantial gap in aural word knowledge in and out of context. It also remains to be seen to what extent tests of spoken vocabulary knowledge can simply be adaptations of existing written tests.

**Speed of Access**

It seems obvious that one characteristic of fluent language use is rapid access to the mental lexicon, so that the user can perform communicative tasks in real time without consciously searching for the necessary vocabulary. However, until recently vocabulary tests have focused on accuracy of response to the test items, without taking into account how quickly the test-taker responds. In this regard L2 vocabulary researchers have approached the investigation of word knowledge quite differently from psycholinguists, for whom reaction time (RT) measures have been a routine tool for many years (see Godfroid, this volume).

An early initiative to incorporate speed of response into an L2 vocabulary test was Laufer and Nation's (2001) study of a computer-based version of the Vocabulary Levels Test. There was significant variation in response times among learners at different levels of proficiency. The researchers also identified an apparent lag in performance whereby the learners responded faster to words at a particular frequency level only after they had demonstrated a good knowledge of that level in terms of accuracy.

Subsequent investigations of response fluency have almost all involved computer-based Yes/No vocabulary tests. Harrington (2006) sought to bridge the gap between the psycholinguistic and SLA perspectives, not only by applying reaction time measures in a Yes/No test but also introducing Segalowitz's (2010) coefficient of variation (CV) as a

measure of how automatically participants responded to the stimulus words. Harrington administered a Yes/No test to participants at three proficiency levels: intermediate ESL, advanced ESL and English as L1. After correction for guessing, the accuracy scores showed consistent increases across the proficiency levels and conversely both the mean reaction time and the mean CV decreased from the intermediate learners through to the native speakers. On the other hand, in a study with a somewhat similar design Miralpeix and Meara (2014) found very little relationship between vocabulary size and either reaction time or CV among students studying English at the University of Barcelona.

More promising evidence of the value of reaction times to evaluate Yes/No test performance was found by Pellicer-Sánchez and Schmitt (2012) in two linked studies. Their basic question was whether RT values could be an alternative to nonwords for adjusting Yes/No scores for overestimation of knowledge. The analysis revealed that the participants responded significantly faster when a Yes response was accurate (as confirmed by a subsequent interview) than when it was inaccurate, and the researchers calculated a threshold reaction time to distinguish the two types of response. This proved to be a better basis for measuring knowledge of the target words than the standard approach of including nonwords and then applying a correction formula. However the participants were highly proficient English users who produced few if any false alarms overall, so it remains to be seen whether these findings can be generalised to less proficient learners with more variable performance on a Y/N test.

The other main relevant research is a series of studies undertaken by Harrington and his associates in various educational contexts, using what they now call the Timed Yes/No test (TYN). In these investigations both accuracy and response speed have been measured. Harrington & Carey (2009) explored the value of a TYN test as a placement measure in a language school in Australia. The TYN results were comparable to those of the existing

school placement test, although the reaction times were a little less effective at discriminating proficiency levels in the school programme.

Since then, Harrington and Roche have conducted several studies (eg, Harrington & Roche 2014, Roche et al. 2016) with students at higher education institutions in Oman, where English has been adopted as the medium of instruction. The studies have looked at the TYN as a measure of readiness for undergraduate study both before and after admission to a degree programme. Generally speaking, the correlations between the TYN results and criterion measures have been quite modest, usually less than .40. Despite the efforts of the researchers to inform the test-takers about the nature of the TYN task, the students have persistently responded Yes to nonwords. According to the researchers, the students' lack of motivation to perform well in a low-stakes test situation was one factor influencing the results. Incidentally, the role of motivation in vocabulary test performance was highlighted by Nation (2007), and there is a substantial body of research on the topic in the educational measurement literature (Wise & Smith 2016).

Drawing on the Oman studies and numerous others, Harrington (2018) has proposed the construct of lexical facility, which is defined conceptually as 'the capacity to recognize words quickly' as a key component of fluent text comprehension, and operationally as a composite measure of accuracy, mean reaction time, and coefficient of variation (CV). He argues that the combination of accuracy and speed is a more sensitive measure of second language vocabulary competence, and more particularly academic language proficiency, than the accuracy scores produced by conventional vocabulary tests. The TYN has its limitations, as Harrington freely acknowledges, but this line of research may stimulate the development of other measures which take account of learners' ability to access their vocabulary knowledge fluently.

Oral vocabulary and speed of access are two dimensions among several that are the focus of the current vibrant upsurge in research on vocabulary testing, as the other chapters in this section of the volume amply demonstrate.

## FURTHER READING

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

This was the first comprehensive account of the assessment of second language vocabulary knowledge.

Milton, J. 2009. *Measuring second language vocabulary acquisition.* Bristol: Multilingual Matters.

This volume draws together research findings by the author along with his colleagues and doctoral students at Swansea University in the UK.

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

Nation, I.S.P., & Webb, S. 2011. *Researching and analysing vocabulary*. Boston, MA: Heinle Cengage Learning.

These two manuals on conducting various types of vocabulary research include extensive discussion and advice on measurement issues.

Read, J. 2013. Research timeline: Second language vocabulary assessment. *Language Teaching, 46 (1)*, 41-52.

This article presents an annotated bibliography of key developments in L2 vocabulary assessment from the 1980s to the present.

## RELATED TOPICS

**REFERENCES**


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

Beglar, D. 2010. A Rasch-based validation of the Vocabulary Size Test. *Language Testing 27(1)*, 101-118.

Cenoz, J., & Gorter, D., eds., 2015. *Multilingual education: Between language learning and translanguaging*. Cambridge: Cambridge University Press.

Cobb, T. 2000. One size fits all? Francophone learners and English vocabulary tests. *Canadian Modern Language Review 57(2)*, 295-324.

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

Du, Y. 2016. *The use of first and second language in Chinese university EFL classrooms*. Singapore: Springer.

Ebel, R.L., & Frisbie, D.A. 1986. *Essentials of educational measurement* 4th ed. Englewood Cliffs, NJ: Prentice-Hall.

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

Fitzpatrick, T., & Barfield, A. eds. 2009. *Lexical processing in second language learners Papers and Perspectives in honour of Paul Meara.* Bristol: Multilingual Matters.

Fountain, R. & Nation, P. 2000. A vocabulary-based graded dictation test. *RELC Journal 31(2)*, 29-44.

Garcia, O. & Li, W., 2013. *Translanguaging: Language, bilingualism and education*. Basingstoke: Palgrave Macmillan.

Gyllstad, H., Valkaite, L., & Schmitt, N. 2015. Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL International Journal of Applied Linguistics 166(2)*, 276–303.

Harrington, M. 2006. The lexical decision task as a measure of L2 lexical proficiency. *EUROSLA Yearbook, 6*, 147-168.

Harrington, M. 2018. *Lexical facility: Size, recognition speed and consistency as dimensions of second language vocabulary knowledge*. Basingstoke: Palgrave Macmillan.

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

Harrington, M., & Roche, T. 2014. Post-enrolment language assessment for identifying at-risk students in in English-as-a-Lingua-Franca university settings. *Journal of English for Academic Purposes 15 (1)*, 37-47.

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

Jordan, E. 2012. Cognates in vocabulary size testing – A distorting influence? *Language Testing in Asia 2(3)*, 5-17.

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

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

Laufer, B., & Goldstein, Z. 2004. Testing vocabulary knowledge: size, strength and computer adaptiveness. *Language Learning 54(3)*, 399-436.

Laufer, B., & Levitsky-Aviad, T. 2018. Loanword proportion in vocabulary size tests: Does it make a difference? *ITL – International Journal of Applied Linguistics, 169(1)*, 94-114.

Laufer, B., & McLean, S. 2016. Loanwords and Vocabulary Size Test scores: A case of different estimates for different L1 learners. *Language Assessment Quarterly 13(3)*, 202-217.

Laufer, B., & Nation, P. 1995. Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics 16*, 307-322.

Laufer, B., & Nation, P. 2001. Passive vocabulary size and the speed of meaning recognition: Are they related? *EUROSLA Yearbook, 1*, 7-28.

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

McLean, S., Kramer, B., & Stewart, J. 2015. An empirical examination of the effect of guessing on Vocabulary Size Test scores. *Vocabulary Learning and Instruction, 4(1)*, 26-35.

Meara, P., & Jones, G. 1988. Vocabulary size as placement indicator. In P. Grunwell, ed., *Applied linguistics in society* (pp. 80-87). London: CILT.

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

Meara, P. & Miralpeix, I. 2016. *Tools for researching vocabulary.* Bristol: Multilingual Matters.

Milton, J., & Hopkins, N. 2006. Comparing phonological and orthographic size: Do vocabulary tests underestimate the knowledge of some learners? *Canadian Modern Language Review 63(2)*, 127-147.

Milton, J. & Riordan, O. 2006. Level and script effects in the phonological and orthographic size of Arabic and Farsi speakers. In P. Davidson, C. Coombe, D. Lloyd & D. Palfreyman, eds., *Teaching and learning vocabulary in another language* (pp. 122-133). Dubai: TESOL Arabia.

Milton, J., Wade, J. & Hopkins, N. 2010. Aural word recognition and oral competence in a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, M. Torreblanca-López, & M.D. López-Jiménez, eds., *Further insights into nonnative vocabulary teaching and learning* (pp. 83-97). Bristol: Multilingual Matters.

Miralpeix, I., & Meara, P. 2014. Knowledge of the written form. In J. Milton & T. Fitzpatrick eds., *Dimensions of vocabulary knowledge* (pp. 30-44). Basingstoke: Palgrave Macmillan.

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

Nation, P. 1983. Testing and teaching vocabulary. *Guidelines 5*, 12-25.

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

Nation, P. 2007. Fundamental issues in modelling and assessing vocabulary knowledge. In H. Daller, J. Milton & J. Treffers-Daller, eds., *Modelling and assessing vocabulary knowledge* (pp. 35-43). Cambridge: Cambridge University Press.

Nation, P. 2012. The Vocabulary Levels Test. Unpublished paper. Retrieved June 30, 2017 from: http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Vocabulary-Size-Test-information-and-specifications.pdf

Nation, I.S.P. 2013. *Learning vocabulary in another language*. 2nd ed. Cambridge: Cambridge University Press.

Nation, I.S.P. 2016. *Making and using word lists for language learning and testing*. Amsterdam: John Benjamins.

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

Nguyen, L.T.C., & Nation, P. 2011. A bilingual vocabulary size test of English for Vietnamese learners. *RELC Journal 42(1)*, 86-99.

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

Read, J. 1988. Measuring the vocabulary knowledge of second language learners. *RELC Journal 19(1)*, 12-25.

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

Read, J. 1998. Validating a test to measure depth of vocabulary knowledge. In A. Kunnan, ed., *Validation in language assessment* (pp. 41-60). Mahwah, NJ: Erlbaum.

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

Revier, R.L. 2009. Evaluating a new test of *whole* English collocations. In A. Barfield and H. Gyllstad, eds., *Researching collocations in another language* (pp. 125-138). Basingstoke: Palgrave Macmillan.

Roche, T, Harrington, M., Sinha, Y., & Denman, C. 2016. Vocabulary recognition skill as a screening tool in English-as-a-Lingua-Franca university settings. In J. Read ed., *Post-admission language assessment of university students* (pp. 159-178). Cham, Switzerland: Springer.

Sasao, Y., & Webb, S. 2017. The Word Part Levels Test. *Language Teaching Research 21(1)*, 12-30.

Sasao, Y., & Webb, S. 2018. The guessing from context test. *ITL – International Journal of Applied Linguistics, 169(1)*, 115-141.

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

Schmitt, N. 2014. Size and depth of vocabulary knowledge: What the research shows. *Language Learning 64(4)*, 913-951.

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

Schmitt, N., Ng, J.W.C., & Garras, J. 2011. The Word Associates Format: Validation evidence. *Language Testing 28(1)*, 105-126.

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)*, 145-171.

Schoonen, R., & Verhallen, M. 2008. The assessment of deep word knowledge in young first and second language learners. *Language Testing 25(2)*, 211-236.

Segalowitz, N. 2010. *Cognitive bases of second language fluency*. New York: Routledge.

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

Stewart, J. & White, D.A. 2011. Estimating guessing effects on the Vocabulary Levels Test for differing degrees of word knowledge. *TESOL Quarterly 45(2)*, 370-380.

Stoeckel, T., Bennett, P., & McLean, S. 2016. Is "I don't know" a viable answer choice on the Vocabulary Size Test? *TESOL Quarterly 50(4)*, 965-975.

Stubbe, R. 2015. Replacing translation tests with Yes/No tests. *Vocabulary Learning and Instruction 4(2)*, 38-48.

Turnbull, M. & Dailey-O'Cain, J., eds., 2009. *First language use in second and foreign language learning*. Bristol: Multilingual Matters, 2009.

van Zeeland, H. 2013. L2 vocabulary knowledge in and out of context: Is it the same for reading and listening? *Australian Review of Applied Linguistics 36(1)*, 52-70.

Webb, S. 2007. The effects of repetition on vocabulary knowledge. *Applied Linguistics 28(1)*, 46-65.

Webb, S., Sasao, Y., & Balance, O. 2017. The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL – International Journal of Applied Linguistics, 168(1)*, 33-69.

Wesche, M.B., & Paribakht, T.S. 1996. Assessing second language vocabulary knowledge: Depth vs. breadth. *Canadian Modern Language Review 53(1)*, 13-39.

Wise, S. L., & Smith, L. F. 2016. The validity of assessment when students don't give good effort. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 204-220). New York: Routledge.

Zhang, X. 2013. The "I don't know" option in the Vocabulary Size Test. *TESOL Quarterly 47(4)*, 790-811.

Zhao, P., & Ji, X. 2016. Validation of the Mandarin version of the Vocabulary Size Test. *RELC Journal*. [OnlineFirst] https://doi.org/10.1177/0033688216639761

[8592 words]

## BIOGRAPHICAL NOTE

John Read is Professor of Applied Language Studies at the University of Auckland, New Zealand. His research and scholarship have focused on the assessment of second language vocabulary knowledge and the testing of English for academic and occupational purposes.