*Language Teaching* (2020), **53**, 109–120
doi:10.1017/S0261444819000326

CAMBRIDGE
UNIVERSITY PRESS

FIRST PERSON SINGULAR

# Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation

Norbert Schmitt[1]*, Paul Nation[2], and Benjamin Kremmel[3]

[1]University of Nottingham, United Kingdom, [2]Victoria University of Wellington, New Zealand and [3]University of Innsbruck, Austria
*Corresponding author. Email: norbert.schmitt@nottingham.ac.uk

**Abstract**
Recently, a large number of vocabulary tests have been made available to language teachers, testers, and researchers.  Unfortunately, most of them have been launched with inadequate validation evidence. The field of language testing has become increasingly more rigorous in the area of test validation, but developers of vocabulary tests have generally not given validation sufficient attention in the past. This paper argues for more rigorous and systematic procedures for test development, starting from a more precise specification of the test's purpose, intended testees and educational context, the particular aspects of vocabulary knowledge which are being measured, and the way in which the test scores should be interpreted. It also calls for greater assessment literacy among vocabulary test developers, and greater support for the end users of the tests, for instance, with the provision of detailed users' manuals. Overall, the authors present what they feel are the minimum requirements for vocabulary test development and validation. They argue that the field should self-police itself more rigorously to ensure that these requirements are met or exceeded, and made explicit for those using vocabulary tests.

## Introduction

It has probably never been easier to design and develop a vocabulary test. However, this was not always the case. As vocabulary re-emerged as an area of interest in the 1980s–1990s (Meara, 1987), a handful of scholars began developing vocabulary tests for teachers and researchers. Early innovators like Paul Nation (*Vocabulary levels test* – 1983) and Paul Meara (*Eurocentres Vocabulary size tests 10KA* – 1990) had to work with manual (and early computerized) frequency counts, and to develop new test formats more-or-less from scratch. Nowadays, a wide range of much more powerful corpora (and the relevant tools to use them) have become readily available, and platforms for designing and delivering tests have never been more affordable and accessible. There is also an unregulated internet where tests can be made public with little or no vetting. In addition, there currently seems to be an unprecedented interest in vocabulary assessment, and this has resulted in a proliferation of vocabulary tests. With the available resources and ever-increasing interest, it seems that today anybody can make a vocabulary test. But that does not mean they should. Testing is a specialized area and requires a great deal of expertise to do well. Just because one can sample words from a corpus and then place those words into an old test format, it does not mean that the resulting 'new' test will be any good. However, we see this 'quick and easy' test development happening all the time.

One could argue that the proliferation of unvalidated tests is no bad thing: end users can try them out and see if they work; let the market handle it. If they don't work, users can switch to other tests. Or the malfunctioning tests can be improved and updated based on ongoing evidence. However, we think these arguments are naive. Most end users lack testing expertise and are not well-equipped to judge the quality of the tests they are using. Also, if a 'famous name' is attached to the test, they may feel reluctant to challenge the test, for who are they to question the experts? Likewise, we note that most tests, once launched, are not revised in any systematic way (or if they are, this is often hardly visible to the users). For example, the *Vocabulary levels test* (Schmitt et al., 2001) is still a well-used standard vocabulary measurement, yet the authors have not updated it at all since it was launched over 17 years ago. The typical practice seems to be to develop a test, get a journal article published on it, and then move on to the next project.

In the early days of the vocabulary reemergence, it was fine for pioneering test developers (including some of the authors of this commentary) to make the best tests they could with the tools and resources they had at hand. But the fields of both vocabulary and language testing have moved on, and we know much more about how to make good assessments. Because of this, we want to argue strongly that vocabulary tests should henceforth be developed in a much more systematic manner, taking advantage of this more highly developed knowledge. In particular, we have a much better idea about how to go about validating new tests to ensure that they actually work in the manner intended. New vocabulary tests should not be launched (on an unsuspecting public) until there is substantial amount of validation evidence supporting them. This will inevitably result in fewer tests being published (which we do not necessarily think a bad thing), but the tests which do pass the more stringent validation requirements will surely be better, and the field will be more justified in having confidence in them. In this commentary, we discuss what we feel are the minimum requirements for vocabulary test development and validation and argue that the field should self-police itself more rigorously to ensure that these requirements are met or exceeded and made explicit for those using such tests.

## Conceptualizing new vocabulary tests

In the Introduction, we mentioned the proliferation of vocabulary tests. This partial list of tests which purportedly measure vocabulary breadth gives some idea of this: *EFL vocabulary tests* (Meara, 1992) and all their language and modality variations (e.g. *AuralLex*, Milton & Hopkins, 2006), several iterations of the *Vocabulary levels test* (VLT) (Nation, 1983; Schmitt et al., 2001), a *New vocabulary levels test* (McLean & Kramer, 2015), an *Updated vocabulary levels test* (Webb, Sasao, & Ballance, 2017), the *Productive vocabulary levels test* (PVLT) (Laufer & Nation, 1999), the *Listening vocabulary levels test* (McLean, Kramer, & Beglar, 2015), the *Lex30* (Meara & Fitzpatrick, 2000), the *Lexical test for advanced learners of English* (Lemhöfer & Broersma, 2012), the revamped *Computer adaptive test of size and strength* (CATSS) (Laufer & Goldstein, 2004; Levitzky-Aviad, Mizrahi, & Laufer, 2014), the increasingly prominent *Vocabulary size test* (VST) (Nation & Beglar, 2007), which is available in multiple variations and bilingual adaptations (e.g. Nguyen & Nation, 2011; Coxhead, Nation, & Sim, 2015).

All these tests claim to measure vocabulary size, but which one to use? The problem is that most of the tests (or rather the publications and literature supporting them) give little help in making a principled decision. Most give rather vague or multi-purpose explanations such as:

- The VLT is designed to give an estimate of vocabulary size for second language (L2) learners of general or academic English (Schmitt et al., 2001, p. 55).
- The VST was developed to provide a reliable, accurate, and comprehensive measure of a learner's vocabulary size from the 1st 1,000 to the 14th 1,000 word families of English (Nation & Beglar, 2007, p. 9). (Note a more detailed description was later put up on Nation's website; see section headed 'Support for end users' below.)
- The CATSS test was designed with a number of possible applications in mind. First, it has the potential to provide researchers and teachers with a diagnostic profile of a learner's present level of vocabulary interlanguage development. Second, it offers a means of measuring the increase in size and strength of L2 vocabulary acquisition as learners' language proficiency develops. Finally, since lexical knowledge has been shown to be related to success in reading, writing, and general language proficiency as well as to academic achievement, the test would allow for more efficient placement and admission in language teaching programmes (Laufer, Elder, Hill, & Congdon, 2004).

There is often no concrete mention in test descriptions about what purpose the resulting test scores should be used for, or the kinds of learners/educational contexts for which the tests might be suitable. The end users are basically left free to decide how to use the tests and to interpret the test scores. In effect, these are 'One Size Fits All' tests, because if there is no stated purpose or context, then presumably any purpose or context is allowed. However, this does not make sense to us. For example, one would expect a vocabulary test used for diagnostic purposes to identify reading vocabulary shortfalls in children to be different from a vocabulary test designed to be a placement test for adult English for academic purposes pre-sessional courses. Thus, the form a vocabulary test takes follows directly from the test's intended purpose/learners/context.

Indeed, current test development and validation theory (e.g. Read, 2000; Chapelle, 2012; Kane, 2012) goes against a 'One Test for Everything' viewpoint. One of the key requirements is to clearly state the purpose of a test, because without an explicit purpose, it is impossible to validate a test (i.e. determine whether the test achieves its purpose or not). This is closely related to score interpretation: the ways the scores will be used in understanding learners' language ability. Vocabulary tests typically give numerical scores (e.g. the percentage of words known at the 2,000 (2 K) frequency level), but without a clearly defined purpose, it is difficult to know what the numbers mean.

Likewise, no vocabulary test (or any language test) can be suitable for all learners in all contexts, so unless the test is appropriate for particular learners and contexts, the test scores can be very misleading. A good example of this is when the VLT was used in the New Zealand school system by a classroom teacher, it gave depressingly low scores for some students. This did not seem to match some teacher's observations of the students' ability, and a study was carried out where the VLT was re-administered one-to-one with a researcher who could encourage the students to answer, and prompt for further information if necessary. The results showed that the one-to-one administration produced much higher vocabulary size estimates than the whole-class administration (Nation, 2007). A follow-up study using the VST (Coxhead & Nation, forthcoming) showed that lower -scoring students simply did not relate to the VST given in a whole-class environment; thus, their vocabulary size was vastly underestimated. Therefore, the test given in a whole-class administration was unsuitable for this particular kind of student, and the misleading poor results could have been quite harmful to their educational prospects.

Therefore, one of the most important improvements we would like to see in vocabulary testing is a better specification of the purpose the test is being developed for. Likewise, it should be specified what type of learners and educational contexts the test has been designed for. These specifications are essential for both developing the test and then for validating whether it meets the specifications.

In addition to purpose/learner/context specifications, the conceptualization of a test should include whether the test is measuring receptive or productive knowledge (or both), and spoken or written knowledge (or both). It should also specify which skill(s) – listening, speaking, reading, and writing – is being focused on, as practitioners/researchers are interested in what learners can do with their vocabulary, rather than what they can merely answer on a test (Schmitt, 2014).

Another important issue is deciding the level of mastery which will be measured. Laufer and Goldstein (2004) found that learners typically knew more words at the 'meaning recognition' level than they did at the 'form recall' level. Thus, tests at these two different levels of mastery will give different results for the same learner. Unless the level of mastery is made clear, it is difficult to interpret the resulting scores sensibly. There has also been a trend towards testing multiple levels to get more informative results (as in the CATSS test), and this should be encouraged. Likewise, measuring non-meaning types of word knowledge (e.g. collocation, derivatives) can give useful information about the learner's depth of knowledge about the words tested (e.g. Webb, 2005).

In sum, test developers need to be very clear why they are making a test (purpose), who it is intended for and in what context, and what aspect(s)/level(s) of vocabulary knowledge they are trying to measure.

## Developing the required expertise

One obvious step in developing the required expertise is doing a thorough critical survey of previous directly-relevant vocabulary research. It must be said, however, that the literature on vocabulary assessment is growing exponentially, and future reviews will need to be more extensive than in the past. The literature on vocabulary in general is bourgeoning, and much of it is showing the complexity of vocabulary knowledge and acquisition (e.g. Webb, 2005; González-Fernández & Schmitt, forthcoming). Therefore, it should not be surprising that testing this complex knowledge will require more sophisticated and precise vocabulary measurement, and more extensive literature surveys will be necessary to provide this. Also, assumptions about foundational conventions like frequency and item types are beginning to be challenged (e.g. Kremmel, 2016; Kremmel & Schmitt, 2016), and test writers will have to stay on top of these developments.

Most previous vocabulary test reports have shown that a VOCABULARY-BASED survey has been carried out, but there has seldom been evidence of a companion review of the LANGUAGE TESTING literature. Vocabulary assessments, albeit having their specific idiosyncrasies, are language assessments. As such, it appears to make sense that the field of vocabulary assessment looks to the field of language assessment, as it provides us with extensive literature on tools and approaches for test development and validation. Looking towards language assessment research, vocabulary researchers and test designers might realize that there are workable frameworks and validation models, even though they have largely ignored them to date. For example, recent argument-based frameworks (e.g. Chapelle, 2012; Kane, 2012) can certainly inform about minimal validation requirements for vocabulary tests, such as the need for test purpose to be specified, as mentioned above. We feel vocabulary testers need to have expertise in both vocabulary and testing. Overall, there needs to be a general improvement in the language assessment literacy of vocabulary researchers (it wouldn't hurt for teachers as well), and expecting a solid understanding of the key principles and concepts of language assessment in general would be a good start (Harding & Kremmel, forthcoming).

A better understanding of both vocabulary and language assessment issues should also enable a greater level of innovation in vocabulary testing. There are endless duplications of existing tests (e.g. the VLT and VST) which might be incremental steps forward, but a greater level of expertise may allow us to break out of old ways of thinking and produce tests that are dramatically better.

## Test development

It is hard to generalize what steps should be taken to improve the stage of actual development of a new test. This will be largely determined by decisions made in the conceptualization stage (e.g. purpose, learner, context, level of knowledge). However, several points are likely to be important in the development of most types of vocabulary test:

- A critical analysis of which item formats can achieve the stated goals.
- An awareness of the problems of particular item formats (e.g. Gyllstad, Vilkaitė, and Schmitt (2015) on multiple-choice items), and the limitations of what they can tell us about learners' ability to use vocabulary.
- The corpus(es) used for making the word lists for the test should be carefully selected to match the purpose of the test. For instance, if the test is supposed to measure oral vocabulary, it probably makes sense to use a spoken corpus. Ideally, the corpus should be available for others to use.
- The procedure for making the word lists should be carefully considered and clearly described, including the unit of counting, what are counted as words and what are not counted as words, the use of range, frequency and dispersion data, and the size of the word lists. Decisions about each of these factors need to be justified.
- The method of sampling from the word lists should be clearly described so that the procedure could be repeated.

- Test developers should consider the advantages of computer technology and the internet in terms of delivery (e.g. tests on *Lextutor*) and computer adaptive testing (e.g. the CATSS; Kremmel, forthcoming).
- Computers and the internet also offer the possibility of innovative use of pictures, sounds, and video clips which might revolutionize vocabulary testing. Although it will require new multimedia skills, the benefits could be dramatic.

## Test validation

We think this is the part of the test development process which has scope for the greatest improvement. Most previous vocabulary tests were supported by very limited (some bordering on non-existent) validation evidence when they were launched. Examples of this include, but are not limited to:

- *Vocabulary levels test (original 1983 version)*: One paragraph mentions trialling, and concludes the test should not be used with learners with Romance L1s (first language).
- *Vocabulary levels test (revised 2001 version)*: One study with 801 learners of mixed L1s, and a number of strands of validation evidence, but no discussion of how the test should be used with different learners or in different contexts.
- *Productive vocabulary levels test*: One study which showed the test could distinguish between learners with different levels of language proficiency, and a second study which discusses the equivalency of the four test versions. But there was no discussion of what the scores meant regarding the ability to use the target words productively.
- *AuralLex*: One study which shows this phonological test is reliable, and correlates moderately with the written *X-Lex* test (Meara & Milton, 2003) on which it is modelled. But there is no discussion of what the scores mean regarding the ability to understand the target words in spoken discourse.
- *Vocabulary size test (original 2007 version)*: The original article contained no validation information, but a follow-up study by Beglar (2010) showed that the different frequency levels decreased in an expected stair-step manner, and had good reliability. However, there was no discussion of what the scores indicated in regard to the ability to employ the target words.

It is important to note that we are not criticizing the earlier test developers for failing to meet current validation standards. Nevertheless, the above descriptions do show that many vocabulary tests have been launched with little evidence of how the tests would behave once out in the world. It is noteworthy that none of the launch publications for the above tests mentioned any research into what the test scores meant regarding what learners could actually do with the target words (other than answer them on the tests!). This reinforces our point that future vocabulary tests should meet a certain minimum level of validation, and give this type of information.

Validation is seen as an ongoing process, and so tests can never be ‘validated’ in a complete and final manner (Fulcher & Davidson, 2007). So what would indicate a reasonable amount of validation evidence which would permit the use of a test with confidence? This is where the narrowing of uses by purpose, learner, and context becomes very useful. The narrower the specifications, the easier it is to determine whether the test works as desired in those limited situations. This obviously entails research focused directly on those limited contexts. Also, without those specifications, it is impossible to explore whether the test works or not.

There is a large literature in test validation (e.g. Kane, 2006; Chapelle, Enright, & Jamieson, 2008; Bachman & Palmer, 2010), but we would argue, at the very least, that the following points need to be part of any adequate validation argument. The first is SCORE INTERPRETATION, i.e. what do the scores mean in real terms? In the words of Messick (1989), a key scholar in the field of validity, we need the ‘integrative evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores’

(p. 13). Similarly, Kane (who has proposed one of the current state-of-the-art conceptualizations of validation) says that 'to validate an interpretation or use of test scores is to evaluate the plausibility of the claims based on the scores' (2013, p. 1). Unfortunately, this evaluation of 'score interpretation' is more often than not lacking in vocabulary (assessment) research. To address this, we feel that performance on a vocabulary test should be related to some kind of language use, such as listening, speaking, reading, or writing. Ideally, there should be some research data to support this, such as comparing test answers to some kind of measure of language performance (e.g. explaining the meanings of words when given the written word form, or answering a comprehension test of a listening passage in which the target words were embedded). Note that while statistical procedures like Rasch modelling (which basically compares all of the items on the test to each other) are very useful for some aspects of validity, they are unlikely to be informative in this case. What is needed is a direct comparison of a test response to an outside criterion measure of how well that target word can be employed by a learner.

A second point is that test developers should be wary when revising older tests. This has been a popular practice, with the VLT, VST, and the *Word associates test* (Read, 1998), among others, being reworked in various ways. This can be perfectly appropriate if the original test was adequately validated, but very few to date have been. There is the risk that a new revision will merely retread a test which has hidden flaws and thus perpetuate those flaws. A good example of this is the VST. For many years, the only validation evidence for it was Beglar's (2010) study of Japanese learners of English. While a good study, it was clearly not enough to 'validate' the VST and show how it behaved for various learners and contexts. For example, one of the things it did not directly investigate was examinee test-taking behaviour of the VST's multiple-choice items. We have always known that multiple-choice items are prone to strategic examinee guessing, and Gyllstad et al. (2015) demonstrated that the multiple-choice VST items typically lead to overestimation of vocabulary size, by as much as 26% at the 9 K frequency level. Clearly this is an issue which needed to be addressed in any revision. Yet, many bilingual versions of the VST were created, all founded on the assumption that the VST was valid based on Beglar's (2010) single study: for example, a Vietnamese version trialled on only 62 students (Nguyen & Nation, 2011), a Persian version piloted on 190 students (Karami, 2012), a Russian version piloted on 121 students (Elgort, 2013), and a Japanese version piloted on 43 students (Derrah & Rowe, 2015). Notably, none of the publications of these bilingual versions make any mention of the multiple-choice overestimation issue. This uncritical view of the original VST is unfortunate because it meant that any potential problems inherent in the original were never investigated and improved upon.

Another problem stemming from an uncritical view of validation is the assumption that the new revised test does not require much further validation evidence. Current validation theory would view any revised version as a new test, which needs to be validated in its own right. It is no good to assume the validity of a test with new items, and potentially different length and format/modality, based only on the behaviour of the original version. In the case of the VST bilingual versions, the changes involved a complete change of language. We know that speakers of various L1s can have quite different behaviour from one another (Dörnyei & Ryan, 2015), so it is unrealistic to assume that the change of language would not be connected to any other change in examinee behaviour. Nevertheless, to our knowledge, no bilingual version has been subject to a comprehensive and rigorous validation procedure.

It may well turn out that some or all of the bilingual VST versions work well, but that needs to be demonstrated. That is what validation is essentially about. Unless a test developer looks at the original test critically, and then carries out substantial validation research on the revised version, there is little assurance that the end users will have a revised test that works for them.

Revision of a test can be a good thing, and hopefully improvement is a result. But the revision process should be documented in some way. The VST has been constantly updated, and the current version is surely better than the original. But none of the updates have been labelled in any way (other than the expansion from a 1 K–14 K version to a 1 K–20 K version), and the current VST is still listed as simply the VST. Users will have no idea which version they are using, short of making a detailed analysis of each individual item on the test. We recommend labelling each new iteration of a test (e.g. Version, 1, 2, 3, or Version 5/25/2012, 11/30/2017), so that practitioners and researchers can instantly see whether they are using the same or different versions of a test.

A third point concerns RELIABILITY, that is, the consistency of the test scores. If a test produces an estimated vocabulary size of 5,000 lemmas for a particular learner today, it should produce a very similar (although probably not exact) score tomorrow for her. Reliability has typically been measured by 'internal consistency' statistics, like Cronbach's alpha, but we feel there are two reasons for not doing this. One, Cronbach's alpha has been criticized for mainly being an indicator of how many items a test contains (Goeman & de Jong, 2018). Vocabulary tests usually have many items, which renders this index less informative. Second, we question the extent to which internal consistency applies to vocabulary tests. When the construct being measured is a single rule or ability (such as 3rd person –*s*, reading ability), it is possible to create tests where all the items address that single construct (e.g. the response to one item targeting a particular reading behaviour might allow us to predict the performance on another item targeting the same behaviour). However, vocabulary is largely item-based learning, and so each item addresses a separate construct, i.e. knowledge of a single lexical item. Thus, just because one lexical item is known (e.g. the word *clever* – #5,028 in the COCA corpus) it does not mean that another will be known (*interactive* – #5,027). This is true even if the words are similar in terms of frequency, form, meaning, or topic area. This causes problems for internal consistency methods, which work by assuming that all test items are measuring the same thing. For vocabulary, in which difficulty depends largely on a learner's idiosyncratic exposure to each word (i.e. some combination of incidental exposure and instruction), it is best to establish reliability using a test-retest procedure (see Schmitt, 2010 for more details). Alternatively, newer approaches, such as the concept of SUMMABILITY (the 'proportion of total test variation that is explained by the sum score' (Goeman & de Jong, 2018, p. 54) might be incorporated into vocabulary validation procedures.

A fourth point concerns piloting and validating the test on the same kind of learner that the test is intended for. Tests are often piloted on university students because they are more readily accessed by academic researchers, and because ethics approval is easier to obtain for that learner group. However, these students may be very different from other learners of different ages, with other levels of proficiency, experience, etc.

As a fifth point, we feel it is important not to launch a new test until a substantial amount of validation evidence has been marshalled. Historically, the field has not been good at collecting validation evidence post-hoc on new tests, and even when additional evidence has been collected, there is no guarantee that the test's original developers will go to the trouble of revising the test based on that evidence (e.g. Schmitt, Schmitt, & Clapham never revised the VLT based on later research and critiques). Also, if there are hidden flaws in the test, many learners may be mis-measured before those flaws come to light in validation studies years later. Therefore, tests need to be validated to the greatest extent possible BEFORE they are launched. A small quiz used in-house probably does not warrant extended validation (though a good knowledge of testing principles always helps in creating such quizzes!). But once a test is disseminated to the public, either through a refereed journal or through an unvetted internet site, a wide range of users may well utilize the test. We know that most vocabulary tests are mainly used in low-stakes contexts, and so the use of unvalidated tests is not a matter of life or death. Nevertheless, it is clear that even low-stakes tests can have effects on learners. Also, these tests are often used in second language acquisition (SLA) research, and therefore impact on theoretical and pedagogical claims. Thus, we feel test developers have an ethical obligation to make sure their tests are well-functioning if they are going to make them generally available.

It is impossible to state a set level of validation which should be required before dissemination. We obviously cannot wait for all aspects of all stages of development to be addressed in full, or we would never be able to launch a vocabulary test. But as a rule of thumb, the more widely used a test is likely to be, the surer we need to be that it works as advertised (i.e. requiring more [and more convincing] validation evidence). This will almost certainly entail more than the traditional single one-off study with a small number of participants. Also, higher-stakes vocabulary tests resulting in stronger claims will require more validation evidence. Certainly, enough evidence needs to be presented so that end users can make informed judgments about whether or not a particular test is fit for their purpose. We feel that, minimally, the various key areas discussed in this commentary should be inspected empirically before tests are launched. It might not be necessary to have each of them addressed in their entirety at the launch of the test, but the field should be given evidence that these aspects were considered, documented, and investigated. And because test validation is always an ongoing process, replicating validation studies in different linguistic, cultural, and proficiency contexts (Porte & McManus, 2019) would almost certainly be informative.

Validation evidence can be assembled in a variety of ways, and we would not like the field to be too prescriptive about how it is done (as long as it is!). However, we encourage vocabulary testers to consider using some of the validation frameworks which have been developed for language testing in general (e.g. Kane, 2006; Chapelle et al., 2008; Bachman & Palmer, 2010). Argument-based approaches to validation start with a clear and explicitly stated purpose and provide structured and comprehensive evidence for justifiable interpretations. This has numerous advantages in guiding both the test development (through the documentation of test specifications, or design blueprints that provide a binding architecture of the test for its purpose) and the test validation process. Vocabulary test development projects such as Voss (2012) and Kremmel (2017) have shown the feasibility of employing such approaches for framing validation evidence for vocabulary tests.

While it is beyond the scope of this commentary to describe in detail what argument-based validation looks like in practice, hopefully the following brief summary gives some sense of the process. As Chapelle (2012) argues, ‘the framework is simple’ (p. 19). ‘First, specify the proposed interpretations and uses of the scores in some detail (i.e. developing an interpretive argument). Second, evaluate the overall plausibility of the proposed interpretations and uses (i.e. compiling a validity argument)’ (Kane, 2012, p. 4). In Chapelle, Enright, and Jamieson’s (2008) framework, for instance, this argument-building involves six steps: (1) domain definition, (2) evaluation, (3) generalization, (4) explanation, (5) extrapolation, (6) utilitization/impact.

The chain of reasoning starts with the DOMAIN DEFINITION. This refers to a careful analysis of the domain from which the vocabulary test items will be sampled. In other words, are the items selected relevant and representative of the domain we want to say something about (e.g. written academic English)? At this stage, a critical survey of previous research in the relevant area should be presented, i.e. we would expect a rationale for the test design (or selection of an existing format) for the purpose at hand. When selecting an existing test/format, this means justifying why the test/format would give a good indication of vocabulary knowledge in the domain in question. When designing a test, it would involve pointing out the weaknesses in existing tests that are too severe for the purpose (or lacking in validation evidence), so that the development of a new test is warranted. This also includes reviewing literature on test design methodology, item types, and use of tests and their results.

Naturally, the clear definition of the test purpose is a prerequisite for this, as the domain definition decisions should already follow logically from that. This involves describing in detail the aspects of vocabulary knowledge we aim to test, as well as the intended appropriate (and also unintended inappropriate) uses of the test and its scores. It also entails whether the test has been designed and validated to be used as a whole, or whether parts of it can be used selectively and independently, and which populations and purposes it has been intended for and tried out on. The evidence required in this domain definition step revolves around item selection (e.g. the representativeness and relevance of the corpus or frequency list that items are sampled from, how well the items selected match with the curriculum), as well as the relevance/effectiveness of the test’s task or item format.

In the EVALUATION step, we move from the analysis of the domain, to the analysis of test scores. Here, evidence should be provided that the test items and scoring procedures are adequate for the intended interpretations. This includes analysing the response behaviours of examinees and running statistical analyses of the test items. This should supply, at the very least, evidence for the careful and sufficient trialling of test items and descriptive statistics of the functioning of the test. Are test items behaving as expected, e.g. in terms of difficulty? For instance, does the difficulty of items sampled from frequency-based lists follow the hypothesized frequency pattern? Is the item quality sufficient in terms of psychometric properties? Are there enough items sampled? This stage clearly has to refer also to the test specifications, and issues around the weighting of individual items or sections in the scoring procedure.

For the GENERALIZATION step, evidence should be provided for the reliability and generalizability of the vocabulary test scores. Test scores should be in line with expectations and should be comparable and consistent across items and administrations. This could, for instance, also involve checking whether test-takers' scores can reliably distinguish between different groups of candidates or learners.

The EXPLANATION step then links the items and scores back to the construct definition. Through evidence from the domain analysis, reference back to the literature, response analysis, and exploring the relationship between the test and other tests of similar constructs or tests of other skill areas, or through examining the internal structure of the test (e.g. factor analyses), we can arrive at support for the meaningfulness of the score interpretation, i.e. that the scores yielded by the test are indeed attributable to the theoretical construct in question. Monitoring test taker behaviour (guessing, test-wiseness, or other factors that might confound interpretations) is also key here. This can be done by employing some kind of criterion measure (e.g. an interview), to check how well test scores actually represent the learner's relevant knowledge.

Next, the EXTRAPOLATION step links the test scores to claims about a candidate's behaviour, knowledge or ability outside of the test situation. At this stage, evidence needs to be provided that the test performance score is a relevant indicator of performance beyond the test setting in the relevant domain. Again, this evidence could come from linking the test scores to scores on other relevant tests. For instance, a vocabulary test claiming to test written receptive form-meaning link knowledge (i.e. the vocabulary knowledge needed for reading) could be administered alongside a reading comprehension test. If the new test of reading vocabulary explained more variance in the reading comprehension test than an existing reading vocabulary test, this would be valuable evidence for the usefulness and validity of the new test.

Finally, the UTILIZATION and IMPACT of the test needs to be examined. This step is often even more neglected than the previous ones that deal with the item properties and qualities, and the provision of statistical information to address them. Utilization is key, however, as we design tests for particular purposes and uses, and unless we are able to show that they are useful for those purpose, all other documentation is relatively limited in its persuasiveness. This means that if a claim is made that this is a diagnostic vocabulary test, then the diagnostic value for learners needs to be empirically demonstrated. If a test is claimed to be a useful tool for learners to identify their gaps in collocation knowledge, then the community should be shown evidence that learners actually find it useful for this purpose. If it is claimed to be a good placement tool, evidence needs to be provided that it actually works in a real-life context. This stage relates again to score interpretation, and more importantly use. It therefore also involves considerations about how scores and uses are communicated to end users, which we will consider in the next section.

## Support for end users

As different stakeholder groups may be interested in the test, different types of supporting materials might have to be offered. A researcher audience will be interested in the technical qualities of the test and the full validity argument (such as suggested in this commentary) will be required. The obvious outlet for this information is research papers, and perhaps on an internet site for the test. On the other hand, learners or teachers will probably favour a straightforward and concise user's manual that outlines how the test is to be used (and not), what its scores mean, and what it can or cannot do in plain language. We feel the simple step of providing user's manuals would go a long way towards promoting effective and appropriate use of vocabulary tests (and the avoidance of misuse). Providing detailed directions on how to administer and interpret the test will help to promote (but not guarantee) the tests being used and interpreted in line with the developer's intentions. One example of this type of support information for the VST can be found on Paul Nation's website (https://www.victoria.ac.nz/lals/about/staff/

publications/paul-nation/Vocabulary-Size-Test-information-and-specifications.pdf), and an example for the CATSS test at (http://catss.cf/index.php). But regardless of how the test is disseminated, the user's manual needs to be prominently available at the same location so that end users will not inadvertently miss it.

We think that the following points would likely be useful for most test users, and so suggest that they be included in the user's manuals:

- The range of possible appropriate uses of the test should be described. It should also be made clear what the test should NOT be used for.
- The way to administer the test should be clearly described, indicating whether it needs to be administered one-to-one and the role of the administrator in this type of delivery, or whether it can be administered to several people at the same time.
- Information about whether the whole test needs to be used, or if parts of it can be used.
- It needs to be made clear whether there is a time limit on administration, or whether learners should be allowed to complete the test in whatever time they need. The typical time taken to complete the test should be indicated.
- Standard instructions for the test should be provided.
- The help that the administrator can give should be described.
- The answers and criteria for marking the test should be provided, including what happens if learners skip an item.
- Guidelines should be provided on interpreting the results. These could include converting the raw score(s) to a vocabulary size estimate, and suggesting what language use learners should be able to achieve given particular vocabulary sizes.
- Performance on the test should be related to some kind of language use, such as listening, speaking, reading, or writing.
- Information on how a typical population of learners (for which the test is intended to be used) performs on the test.

Another way to support end users is for test developers to monitor the use and performance of the test. This can be done by developing a feedback mechanism. Although this has not been a feature of previous vocabulary tests, today's technology makes this feasible. For example, the test developer could ask end users to send in their anonymized test results to help improve the test. It may be even more straightforward if the test is internet-based. Free use of the test could be made contingent on allowing the test developer to retain the test scores for future test revision. This data could also be linked with impact studies into test effectiveness in washback studies.

## Conclusion

In recent times, vocabulary tests have proliferated, but we feel that the validation aspect has not kept pace with this surge. We feel strongly that, as academics and testers, we have the moral duty to produce the very best tests that we can because they have real effects on human examinees. Only by carrying out rigorous validation procedures can we be confident that we are meeting our duty of care to these examinees. The stances we take in this commentary are the result of numerous experiences/discussions that we have had over the course of our intensive and/or long engagement with issues pertaining to vocabulary assessment, and our own vocabulary test development projects. While you may disagree with some of the specifics, we hope you are convinced of the need for a generally greater rigour in vocabulary test development.

If a consensus for greater rigour can be achieved, then it is up to the field to enforce higher validation standards, such as by requiring better validation evidence as journal reviewers, critiquing validation arguments more carefully when reviewing tests, and by not launching new tests until a more substantial amount of validation evidence has been amassed. The eventual result can only be a higher standard of vocabulary test in the future.

## References

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

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

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple…. *Language Testing*, *29*(1), 19–27.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.

Coxhead, A. & Nation, P. (forthcoming). Group vs. individual: How administration procedure can affect vocabulary test scores.

Coxhead, A., Nation, P., & Sim, D. (2015). Vocabulary size and native speaker secondary school students. *New Zealand Journal of Educational Studies*, *50*(1), doi: 10.1007/s40841-015-0002-3

Derrah, R & D.E. Rowe (2015). Validating the Japanese bilingual version of the Vocabulary Size Test. *International Journal of Languages, Literature and Linguistics*, *1*(2), 131–135.

Dörnyei, Z., & Ryan, S. (2015). *The psychology of the language learner revisited*. New York: Routledge.

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the vocabulary size test. *Language Testing*, *30*(2), 253–272.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. New York: Routledge.

Goeman, J. J., & de Jong, N. H. (2018). How well does the sum score summarize the test? Summability as a measure of internal consistency. *Educational Measurement: Issues and Practice*, *37*(2), 54–63.

González Fernández, B., & Schmitt, N. (in press). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*.

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

Harding, L., & Kremmel, B. (in press). SLA researcher assessment literacy. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing*. New York: Routledge.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, *29*(1), 3–17.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

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

Kremmel, B. (2016). Word families and frequency bands in vocabulary tests: Challenging conventions. *TESOL Quarterly*, *50*(4), 976–987.

Kremmel, B. (forthcoming). Developing a computer adaptive test: Which algorithm works best?

Kremmel, B. (2017). *Development and initial validation of a diagnostic computer-adaptive profiler of vocabulary knowledge* (Unpublished Ph.D. thesis). University of Nottingham.

Kremmel, B., & Schmitt, N. (2016). Interpreting vocabulary test scores: What do various item formats tell us about learners' ability to employ words? *Language Assessment Quarterly*, *13*(4), 377–392.

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *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., & Nation, P. (1999). A vocabulary-size test of controlled productive vocabulary. *Language Testing*, *16*(1), 33–51.

Levitzky-Aviad, T., Mizrahi, L., & Laufer, B. (2014). A new test of active vocabulary size. EUROSLA 24, book of abstracts, p. 48.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*(2), 325–343.

McLean, S., & Kramer, B. (2015). The creation of a New Vocabulary Levels Test. *Shiken*, *19*(2), 1–11.

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

Meara, P. (1987). *Vocabulary in a second language*, *Vol. 2*. Specialised Bibliography 4. London: CILT.

Meara, P. (1990). Some notes on the Eurocentres vocabulary tests. Retrieved from http://www.lognostics.co.uk/vlibrary/meara1990b.pdf

Meara, P. (1992). *EFL vocabulary tests*. University College, Swansea: Centre for Applied Language Studies.

Meara, P., & Fitzpatrick, T. (2000). Lex 30: An improved method of assessing productive vocabulary in an L2. *System*, *28*, 19–30.

Meara, P., & Milton, J. (2003). *X_Lex, The Swansea Levels Test*. Newbury: Express.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–104). New York: Macmillan.

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners? *Canadian Modern Language Review*, *63*(1), 127–147.

Nation, I. S. 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 University Press.

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

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

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

Porte, G., & McManus, K. (2019). *Doing replication research in applied linguistics*. New York: Routledge.

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: Lawrence Erlbaum.

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

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., 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.

Voss, E. (2012). *A validity argument for score meaning of a computer-based ESL academic collocational ability test based on a corpus-driven approach to test design* (Unpublished Ph.D. thesis). Iowa State University.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, *27*, 33–52.

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL – International Journal of Applied Linguistics*, *168*(1), 34–70.

**Norbert Schmitt** is Professor of Applied Linguistics at the University of Nottingham. He is interested in all aspects of second language vocabulary. He has published eight books, and over 50 journal articles and 23 book chapters on various vocabulary topics. A special interest is vocabulary measurement, having co-revised the second version of the *Vocabulary Levels Test* (Schmitt, Schmitt, & Clapham, 2001), and having worked with various other lexically-focused measurements ranging from word association results to eye-tracking data. He has sat on the editorial board of *Language Testing* and is a former member of the TOEFL Committee of Examiners. His personal website (www.norbertschmitt.co.uk) gives more information about his research, and also provides a wealth of vocabulary resources for research and teaching.

**Paul Nation** is Emeritus Professor of Applied Linguistics in the School of Linguistics and Applied Language Studies at Victoria University of Wellington, New Zealand. He has now retired from teaching and supervision. His books on vocabulary include *Teaching and learning vocabulary* (1990) and *Researching and analysing vocabulary* (2011) (with Stuart Webb) both from Heinle Cengage Learning., His latest books on vocabulary are *Learning vocabulary in another language* (second edition 2013) published by Cambridge University Press, *Making and using word lists* (2016) from John Benjamins, and *How vocabulary is learned* (2017) (with Stuart Webb). His website contains many resources for teachers and researchers.

**Benjamin Kremmel** is University Assistant at the University of Innsbruck, Austria, where he is heading the Language Testing Research Group Innsbruck. He holds an M.A. in Language Testing from Lancaster University, UK, and a Ph.D. in Applied Linguistics from the University of Nottingham, UK. His research interests are in assessment of L2 lexical knowledge and L2 reading skills, as well as in language assessment literacy. He has published in *Language Testing*, *Language Assessment Quarterly*, *Applied Linguistics*, and *TESOL Quarterly*. Benjamin is the winner of the 2013 Caroline Clapham IELTS Masters Award and the recipient of the 2015 Robert Lado Memorial Award. He serves on the Editorial Boards of *Language Testing*, *Language Assessment Quarterly*, and *ITL International Journal of Applied Linguistics*.

**Cite this article:** Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, *53*(1), 109–120. https://doi.org/10.1017/S0261444819000326