Assessing vocabulary size through multiple-choice formats

Issues with guessing and sampling rates

Henrik Gyllstad, Laura Vilkaite and Norbert Schmitt
Lund University / University of Nottingham

In most tests of vocabulary size, knowledge is assessed through multiple-choice formats. Despite advantages such as ease of scoring, multiple-choice tests (MCT) are accompanied with problems. One of the more central issues has to do with guessing and the presence of other construct-irrelevant strategies that can lead to overestimation of scores. A further challenge when designing vocabulary size tests is that of sampling rate. How many words constitute a representative sample of the underlying population of words that the test is intended to measure? This paper addresses these two issues through a case study based on data from a recent and increasingly used MCT of vocabulary size: the *Vocabulary Size Test*. Using a criterion-related validity approach, our results show that for multiple-choice items sampled from this test, there is a discrepancy between the test scores and the scores obtained from the criterion measure, and that a higher sampling rate would be needed in order to better represent knowledge of the underlying population of words. We offer two main interpretations of these results, and discuss their implications for the construction and use of vocabulary size tests.

Keywords: vocabulary size, multiple-choice test, guessing, sampling rate, assessment, validation, testing, criterion-related validity

Introduction

Vocabulary size is typically defined as the number of words in a language for which an individual has at least a basic form-meaning mapping knowledge (see e.g. Meara, 1996). From a learning point of view, recent research has shown that relatively large vocabulary sizes are necessary to operate successfully in English (e.g. Nation, 2006; Schmitt & Schmitt, 2012) and this implies the need to assess

International Journal of Applied Linguistics 166:2 (2015), **276–303**. DOI 10.1075/itl.166.2.04gyl
ISSN 0019–0829 / E-ISSN 1783–1490

whether learners have acquired (or are in the process of acquiring) the required vocabulary. Vocabulary size is also often used as a proxy for general proficiency in language acquisition research, as vocabulary size scores have been shown to correlate highly with scores on general proficiency tests (Alderson, 2005). With this comes a need to measure vocabulary sizes in a reliable and valid way.

There are two issues that are central in this regard. The first has to do with how the vocabulary knowledge is assessed. The most influential and commonly used format is receptive multiple-choice. Despite its popularity, there are some well-known problems with this format, such as potential overestimation of scores through guessing. The second issue has to do with the sampling of words from the language for inclusion in the test. Because there are usually too many words to take on single test, vocabulary size is typically assessed through a smaller sample of the total population of words. After learners take the test, an extrapolation is made about their vocabulary size based on their scores of that sample. The question is how many words are sufficient to give an accurate estimate, i.e. what is an acceptable sampling rate? This paper reviews these two issues and reports on a case study based on data from administrations of one multiple-choice test, the Vocabulary Size Test (VST) (Nation & Beglar, 2007).

Multiple-choice formats and guessing

A number of influential tests currently exist for vocabulary size measurement in English: the *Vocabulary Levels Test* (VLT) (Nation, 1990; Schmitt, Schmitt, & Clapham, 2001), the *Vocabulary Size Test* (VST) (Nation & Beglar, 2007), the *CATSS* (Laufer & Goldstein, 2004), and checklist (also known as the Yes/No test) tests (Meara & Buxton, 1987; Pellicer-Sánchez & Schmitt, 2012). (See Read (2000) and Schmitt (2010) for more detailed discussion of these and other vocabulary tests.). With the exception of the checklist test, which employs a self-report binary-choice format, the above tests are all, partially or fully, multiple-choice tests (MCT).[1] A vocabulary MCT item consists of an item stem with a target word and set of response options, typically three or more with one keyed as the acceptable answer and the remainder, the distractors, as unacceptable answers. MCTs have a long and time-honoured history in language assessment in general. As early as in the 1920s, Wood (1928), calling them "new-type" tests, argued for their use

1. The VLT uses a matching format, but is still essentially multiple-choice in nature, with each stem having six possible options. Also, it is not strictly a vocabulary size test, but is often used as such (see e.g. Schmitt & Meara, 1997; Stenius Staehr, 2009).

in modern foreign language assessment, mainly due to lower costs for scoring in large-scale administrations.

Although there are some obvious advantages with vocabulary multiple-choice formats, such as their widespread familiarity, ease of scoring, and the fact that a large number of words can be tested in a short period of time, there are also some clear disadvantages. Wesche and Paribakht (1996) list no fewer than six points of criticism. Among these, they claim that test-takers may arrive at a correct answer through a process of elimination, with a standard four-alternative item allowing for a 25 per cent chance of guessing the correct answer, and that items may test knowledge of the distractors rather than a more exact meaning of the target word. It is clear that a correct answer to a MCT vocabulary item can sometimes be achieved through either elimination of distractors rather knowledge of the target word, or through the application of guessing.

In studies investigating the validity and reliability of the VLT, for example, factors other than knowledge of the target word have been found to be at play in the test-taking process. Even though the VLT format was designed to minimize guessing, Stewart & White (2011), using a formula based on elementary probability theory, found that a six-choice format like the VLT is likely to lead to an average score increase of 16.7% due to guessing for most levels of ability. Furthermore, Kamimoto (2008), employing think-aloud protocol methodology, found that lower-proficiency test-takers' mean scores on the VLT were inflated by as much as 45% on the 3K level, i.e. cases where no knowledge of the meaning(s) of the target word was observed, but where the test-takers still chose the correct alternative.

The literature also points to problems with other vocabulary MCTs when it comes to overestimation of vocabulary size, for example, the VST. It is a 140-item test measuring word knowledge at 14 frequency levels (1–14K), with items that consist of the prototypical multiple choice (M-C) format. Target words are placed in a non-defining minimal context stem (see Figure 1), with four definition options that often share some semantic features. Stewart (2014), drawing on computer simulations, argues that the score inflation for a four-choice vocabulary format like the VST could be as high as 25% for most ability levels.

1. soldier: He is a **soldier**.
   a. person in a business
   b. student
   c. person who uses metal
   d. person in the army

Figure 1. An example of a VST item.

On the whole then, vocabulary MCTs are widely used, but there is little hard empirical evidence beyond computer simulations of whether these tests work well. The computer simulations suggest problems of overestimation, but how do the tests function in the real world? That is, do test-takers 'know' the words they are credited with knowing in these tests? There is a need to go beyond computer simulations and explore this question with live examinees.

Sampling issues in vocabulary size testing

In most educational settings, an individual's vocabulary size is generally too large to test all the words separately; therefore sampling becomes an important issue in vocabulary testing. In order to avoid test fatigue with test-takers, test designers opt for shorter rather than longer tests, and aim for designs that do not require extensive resources for administration and scoring. Thus practicality largely determines the number of items that can be included in a test, out of the thousands that are possible. Nation (1993: 36) has advised to: "[c]hoose a sample that is large enough to allow an estimate of vocabulary size that can be given with a reasonable degree of confidence". However, what a 'large enough' sample means remains to be decided by test-designers and there are no hard-and-fast rules for how many items are adequate.

From a theoretical perspective, there are two distinct approaches to sampling test items. Classical Test Theory (CTT) implies that the more items are included in the test, the more valid the result is likely to be. Item Response Theory (IRT) postulates that this is not necessarily the case and shorter tests can work as well as longer tests if the items are selected properly, based on certain statistical models (Embertson, 1996). It remains debatable which of these two approaches works better for vocabulary testing, because vocabulary (item-based) tests may not behave in the same way as proficiency or grammar (system-based) tests. Language tests usually have more than one item, designed to evaluate the same construct, while in vocabulary tests each word could be argued to be an individual construct and knowing one word in a certain frequency band does not necessarily entail knowing the other words in that frequency band. Therefore, because of the specific nature of lexical knowledge, the IRT approach might be potentially problematic for vocabulary size sampling procedures. Thus it is not surprising that the most influential vocabulary size tests are sampled in line with CTT, with practicality issues in mind. Vocabulary tests like the Yes/No test (Meara & Buxton, 1987; Meara, 1992), which requires test-takers to simply indicate if they know the word or not, can include more items (e.g. 40 items per frequency band of 1,000 words), while

tests requiring more complex and time-consuming tasks have to be kept shorter. For example, the VLT tests 30 items per frequency band, and the VST only 10.

In general, the important empirical question is how many items are appropriate to represent an underlying population of words in a frequency band. We will address this question in a case study by looking at one MCT which is already in use: the VST.

In the VST with ten items representing each frequency level in the test (1,000 word families), each target word represents 100 words in the band (i.e. a 1:100 ratio). With such a small number of target items, each one becomes critical, with each word's characteristics (e.g. cognate or false friend or not) and each test item's efficacy (strong or weak item) having a disproportionate effect on the overall vocabulary size estimate. With more target items per band, the effect of any potentially problematic item is lessened. On the other hand, using a Rasch analysis, Beglar (2010) found that test forms with as few as five items per level (a 1:200 ratio) can yield adequate reliability and item separation indices, which has led to creating VST test versions of 5 items per frequency band. However, a probabilistic Rasch approach cannot directly demonstrate how performance on the test items relate to knowledge of the overall population (i.e., all of the words in the relevant 1,000-word frequency band), because it does not employ an external criterion of this knowledge to feed into the analysis. Thus the number of items per frequency level required to describe that level with confidence remains to be empirically determined.

Summing up, the above review has identified potential problems to do with overestimation in vocabulary size MCTs. However, very little hard empirical evidence exists when it comes to investigating whether test-takers' scores on a test reflect demonstrable knowledge. Furthermore, the item-based nature of vocabulary raises questions to do with suitable sampling rates. The question is how many multiple-choice items can adequately reflect test-takers' knowledge of the relevant frequency band. This study sets out to empirically investigate these two issues, with the following research questions:

RQ1. Do scores from the investigated vocabulary size MCT match, underestimate, or overestimate test-takers' demonstrable word knowledge?

RQ2. Is the sampling rate of the investigated vocabulary size MCT sufficient to represent test-takers' knowledge of a relevant frequency band in a valid way? If not, what would a more appropriate sampling rate be?

The case study

Methods

We investigated the RQs using data collected through a new vocabulary test which is gaining popularity, the *Vocabulary Size Test* (VST) (Nation & Beglar, 2007). The rationale for using the VST in our case study is not only that it is a new, increasingly used test, and that it is beginning to be used in research studies as a primary measure (Elgort 2011; Uden, Schmitt, & Schmitt, 2014), but principally that it uses the typical 4-option multiple-choice format. The test was briefly described in the literature review, but some additional information is afforded here. The test was developed by Paul Nation and it first appeared in *The Language Teacher* (Nation & Beglar, 2007) and has been reproduced in books (Nation & Gu, 2007; Schmitt, 2010) and on several websites.[2] At present, the VST exists in two English monolingual versions: the original 1–14,000 word version (14K) and a more recent 1–20,000 word version (20K). This study deals with the former, as it is the version most widely available. The 14K test consists of 140 items, divided into 14 sections of 10 items each. Each section in the test corresponds to an underlying 1,000-word frequency band in a 14,000 word frequency list of English word families developed from the British National Corpus (BNC). The VST was developed to "provide a reliable, accurate, and comprehensive measure of a learner's vocabulary size from the 1st 1000 to the 14th 1000 word families of English" (Nation & Beglar, 2007: 9), with its intended use being mainly pedagogical in nature, namely to determine whether learners have enough vocabulary to read in English: "Users of the test need to be clear what the test is measuring and not measuring. It is measuring written receptive vocabulary knowledge, that is, the vocabulary knowledge required for reading." (Nation, 2012, no page number).

At the time of writing, there has been limited investigation into the characteristics of the VST. As examples of what has been done, Beglar (2010), using IRT techniques, found that the test had some good technical characteristics in terms of adequate reliability and item separation indices, but that some items seemingly needed revision. Gyllstad (2012) found that participants generally scored better at the higher frequency bands and poorer at the lower frequency bands (as expected, and being evidence of construct validity), but that there were many exceptions to this general trend. Just like Beglar, he also found that some items were in need of revision. Overall, then, in the spirit of validation as a continuing process (Messick, 1995), further validation is called for.

---

2. Paul Nation's (https://www.victoria.ac.nz/lals/about/staff/paul-nation); Tom Cobb's *Lextutor* <http://www.lextutor.ca>; <http://my.vocabularysize.com/select/test>

In order to be able to address our first RQ, we needed to have an additional measure of the vocabulary knowledge, to which we could compare test-takers' performance on the MCT to see if they actually knew the MCT words they were credited with knowing. In more technical terms, this implies a criterion-related approach (Bachman, 1990; Weir, 2005), a common method for exploring how well test scores reflect knowledge of an underlying construct. A key aspect of validity is the inferences or interpretations which can be reasonably drawn from the test scores. Thus we used score interpretation as guidance in selecting our criterion measure. Since the VST MCT is intended to be interpreted as a measure of vocabulary which can be utilized in reading, the criterion test format must tap into a degree of mastery which will allow this. Reading requires a 'meaning recall' degree of mastery (i.e. the ability to recall the meaning of a word when its orthographic form is read), and it is an interesting question which type of concurrent measure would best capture this mastery.

The best measure would be to see if the target words could be quickly and automatically recognized and understood in actual reading texts, without the need to stop and inference from context. (The VST is a vocabulary knowledge test, not a test of inferencing ability). But this is difficult to operationalize, with issues such as what qualifies as a natural reading text, and how to construct one long text in which a large number of target words could be naturally embedded. Even if such a text could be developed, the problem remains of how to probe the knowledge of the target words — after reading the whole text or directly after the participants encounter each word. Furthermore, it is not possible to use any kind of MCT as the criterion measure, as this does match the level of vocabulary knowledge required for reading. That is, when reading a text and coming across a word like *porridge*, there are no options to choose from: a. a kind of dog, b. a place to put money, c. something to eat at breakfast. Learners need to know and be able to recall a word's meaning for it to facilitate fluent reading.

Schmitt (2010) argues that perhaps the best way of demonstrating vocabulary meaning recall knowledge is "through interactive face-to-face interviews where the interviewer can probe the examinees [sic] lexical knowledge in detail and come to a very confident determination of this knowledge" (p. 182). We ultimately decided to use this method in conjunction with short, written non-defining sentence contexts containing the target words. Thus, to obtain a good indication of our participants' knowledge of the target words, because of the difficulties involved with a reading passage approach, and because of the inappropriacy of M-C criterion formats, an interview technique was adopted. This technique allowed us to measure participants' knowledge at the meaning recall level, which while being a higher level of knowledge than required for meaning recognition, is the level of mastery needed for reading. It has also previously been successfully used

in validation studies of the Vocabulary Levels Test (Schmitt, Schmitt, & Clapham, 2001) and the Word Associates Format (Read, 1998; Schmitt, Ng, & Garras, 2011).

As to the second RQ dealing with sampling rate, we decided to create a set of additional MCT items. This was done to enable us to tease out how test-taker performance on the original 10 items of the test relate to the performance on the criterion measure, and compare this to how test-takers' performance on an increased number of items relates to the performance on the criterion measure. This made it possible to see how an increase of items from 10 to 15, 20, 25, or 30 items would affect how valid the score on the MCT is compared to the demonstrated knowledge of the words in the frequency band, which the MCT sections are intended to represent.

Sampling items from the entire VST in the study would have been impractical when using our time-intensive interview approach, so we focused on three test levels: higher-frequency (3K), mid-frequency (6K), and lower-frequency (9K) (Schmitt & Schmitt, 2012). This spread of frequency bands should provide an indication of how test-takers perform across a range of frequencies, up to the 9K level needed in order to read a wide range of authentic English texts (Nation, 2006).

Materials

Two principal test instruments were created for the study. Firstly, a written pencil-and-paper test of MCT items was compiled. This was done by taking the original 10 VST items from the 3K, 6K and 9K sections of the test, and adding an additional 20 items per section created specifically for these bands, closely following the procedures and specifications reported in Nation & Beglar (2007) and Nation (2012), and using the same frequency lists as those used for the original test. The new items were piloted with five native speakers of English to assure that they could not guess the correct answer when seeing only the four options without the stem, but that they could correctly answer all of the complete items (stem + options). The resulting 30 items per level were inserted into a test booklet, thus 30 3K items, 30 6K items and 30 9K items.

We then created the second principal test instrument: the criterion measure. We wished to sample from the 1,000 words in the three target frequency levels at a rate which would provide a more complete sample and which would lead to a high level of confidence. There is no agreement about what sampling rate is sufficient, although a classical test theory perspective would suggest that more is better. Therefore, we consulted several testing specialists, and the consensus was that a 1:10 sample rate would satisfy them as a reasonable criterion in this study to compare the MCT results to (i.e. 100 interview measure items). The interview measure thus covered the 30 target words given on the expanded VST, plus another 70

words randomly sampled from the relevant frequency bands. The sampled target words were put in a minimal non-defining context to make the interview items as similar to the test items as possible:

Triangle: He drew a **triangle**.

The three sets of interview measure items were piloted with five native speakers of English using the same procedure as with the VST test items to ensure that the interview items were not guessable from the sentence context.

Participants

Our methodology of comparing correct/incorrect responses to the MCT items with an in-depth oral probing of the same target words necessitated having a situation where participants knew some, but not all, of the target words. If they knew all of the target words, we could not gather information about their test-taking behaviour when they did not know the words (i.e. test-taking strategies). Conversely, if they did not know any of the words, we could not investigate their test-taking behaviour when they did know the target words. We therefore needed population groups where we would get this 'mixed' (some known and some unknown) response pattern for our three chosen frequency levels. After a series of pilots, it was found that an available Lithuanian population was suitable for the 3K level, and a Swedish population was suitable for the 9K level. Furthermore, we were able to find two populations for which the 6K level was suitable, a UK-based mixed-L1 participant group and the aforementioned Swedish population. This allowed an overlap for that frequency band with two participant groups in two different locations being tested on the same material. In total, 141 participants took part in the study. The overall design of the study in terms of participant groups and test sections used is shown in Figure 2.

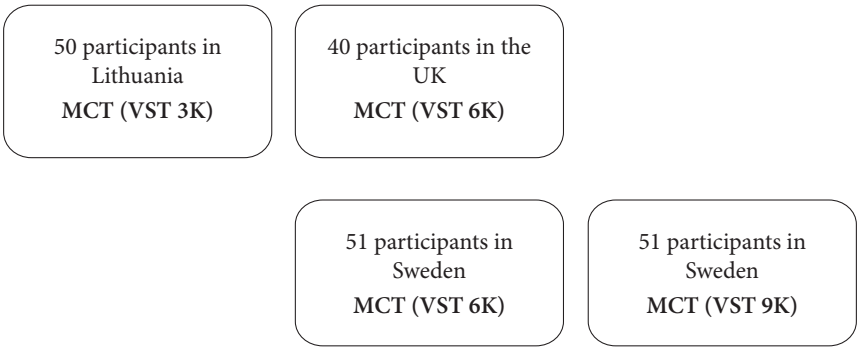

Figure 2. The participant groups of the study and the test sections administered

3K participants in Lithuania: 50 Lithuanians (22 males, 28 females) from various proficiency levels were recruited. Their mean age was 30.5 (SD = 12.21; range: 38 [18, 55]), and while their educational background and proficiency varied considerably, they had all studied English either at school and/or in various language courses. They were either university students (BA students N = 14, MA students N = 10) or working professionals (N = 26).

6K participants in the UK: 40 students at a British university (12 males, 28 females) voluntarily participated in the study. They were recruited from the university's MA in Applied Linguistics program, as well as from various student societies joined by international students. Their mean age was 21.4 (SD = 3.3; range 18 [18, 35]), and they were spread across 19 different L1s, with the largest groupings being Chinese (n = 8), Arabic (n = 6) and Kurdish (n = 4). English was a second language for all of them. They had all met the minimum university entrance requirement of an IELTS 6.0, TOEFL iBT 79, or PTE Academic 55, although many were considerably above this proficiency level.

6K and 9K participants in Sweden: 51 students at a Swedish university (25 males, 26 females), volunteered when taking a module on English vocabulary learning. Their mean age was 24.2 (SD 6.7; range: 22 [18, 39]) and they had just started their first year of study on a five-year teacher training programme. In terms of proficiency level, the grade prerequisites for admission to the programme correspond to a minimum CEFR level of B2. The same individuals took the 6K and 9K tests.

Procedure

Table 1 illustrates the study procedures, and the data were gathered as follows. First, the pencil-and-paper 30-item MCT was administered (for the Swedish groups both the 6K and 9K test booklets were given at the same time). For most participants, 5–10 minutes was enough to sit the 30 items. As the second step, the participants took part in the oral interview.

No explicit link between the pencil-and-paper test and the interview was mentioned. The interview comprising 100 words took on average 25–30 minutes to carry out (50–70 minutes for the Swedish participants for two bands, including a short break in the middle). No fatigue effects were observed. Each interviewee was told that the aim of the interview was to find out whether they knew the meaning of a list of English words. They were asked to look at the prepared list of words and describe the meaning of those words, one by one. Each word was presented as in the MCT item, but without the four alternative options:

peacock: I saw a **peacock**.

Table 1. Study procedure.

Data on the 3K level **Location: Lithuania** **N = 50**	**Data on the 6K level** **Location: the UK** **N = 40**	**Data on the 6K and 9K levels** **Location: Sweden** **N = 51**
Instructions	**Instructions**	**Instructions**
Paper-and-pencil MCT booklet 30 items	**Paper-and-pencil MCT booklet** 30 items	**Paper-and-pencil MCT booklet** 2 x 30 items
Vocabulary interview measure 100 items	**Vocabulary interview measure** 100 items	**Vocabulary interview measure** 2 x 100 items
Retrospection interview		

It was made clear that the non-defining context only illustrated *one* use of potentially many for the word form, and that any meaning sense of a polysemous word would be accepted. Participants were told that they were expected to demonstrate meaning knowledge of the target words in any way that was comfortable for them (e.g., L1 translation equivalent or definition, L2 synonym or definition, picture, gesture). The interviewing researchers had a comprehensive catalogue of L1 translation equivalents and L2 synonyms and definitions available for all the target words during the interview.

In terms of deciding whether an interviewee knew the meaning of a word or not, the following criteria were used. For a target word like *peacock* (6K), if encountering this word in a text, we arguably would like the word form to evoke the concept 'peacock', i.e. some kind of semantic, prototypical representation of the animal. In the interview, it was thus not enough to say 'an animal' or 'some kind of bird', since these were considered merely very broad classifications of the noun in question. Interviewees who answered 'bird' were asked to supply further information to show that they knew something more, for example characteristics like size and/or colour, to distinguish *peacock* from other birds. Likewise, for a target word like *artillery* (6K), only saying that the word to them had military connotations was insufficient; they were expected to supply some more precise meaning information, e.g. 'big, heavy guns'. This means that a general classification plus some kind of additional meaning feature was generally required. Moreover, if the target word was a cognate, then just supplying the L1 word was not considered enough and interviewees were asked what the meaning of the L1 word was to them, in order to rule out that their knowledge was exclusively form-based, devoid of meaning. However, we want to emphasize that a very detailed meaning knowledge of the target words was of course *not* a requirement, and the amount of instances where participants were asked to supply more information was in reality

very small. Surely, only ornithologists and military experts can supply precise and meticulous descriptions of concepts like 'peacock' and 'artillery'.

Finally, in order to gain insights into test-taking behaviour, twenty participants took part in a retrospective interview based on the original ten test items in the 3K band. These participants were selected from the 3K participant group to represent a range of proficiency levels: their vocabulary interview measure scores (Max = 100) ranged from 13 to 90, with a mean of 67. The retrospective interviews were carried out immediately after the vocabulary interview part of the study. The participants were not informed of these interviews beforehand, in order not to affect their test-taking behaviour in any way. They were presented with their tests in order to be able to go through the results and explain their choices. The participants showed little difficulty in remembering the reasons for choosing their answers, and because the interviews were conducted in Lithuanian by the second author, no language-related difficulties in formulating their answers were reported by the participants.

Results

Reliability indices for the test instruments used in the study are shown in Table 2. The reliability indices (Kuder Richardson 21) of the 30-item MCT booklets were acceptable to high (.68–.87), and the 100-item interview produced very high reliability figures (.92–.97).

Descriptive statistics for the paper-and-pencil MCT and the interview criterion measure are provided in Table 3. Results of the test and interview measures confirm that our selection of participant groups did produce the desired range of scores in the three frequency bands, e.g. a mean of 65–69% correct responses on the original 10 VST items. Furthermore, the table shows that the difficulty of the additional items (11–30) matches the difficulty of the original items (1–10) rather well. It is important to note here, though, that the values in Table 3 do not

Table 2. Reliability indices (Kuder Richardson 21) for the MCT test booklet and the interview measure administrations.

	3K Test administration	6K Test administration A	6K Test administration B	9K Test administration
MCT booklet 30 items	.85	.87	.70	.68
Interview criterion measure 100 items	.97	.97	.92	.93

in themselves indicate either under- or overestimation.[3] Subsequent analyses will be needed to address this issue.

RQ1. Do scores from the MCT match, underestimate, or overestimate test-takers' demonstrable word knowledge?

Research Question 1 was addressed through three analyses. Firstly, participants' scores on the original 10 MCT items were compared with their knowledge of the same items in the interview.

Table 3. Descriptive statistics for the MCT test booklet and the interview measure administrations.

MCT Test booklet 30 items	**3K Test administration Lithuania N = 50**			**6K Test administration A United Kingdom N = 40**			**6K Test administration B Sweden N = 51**			**9K Test administration Sweden N = 51**		
	Mean	**SD**	**PC**	**Mean**	**SD**	**PC**	**Mean**	**SD**	**PC**	**Mean**	**SD**	**PC**
10 items	6.88	2.22	.69	6.48	1.89	.65	6.90	1.40	.69	6.49	1.58	.65
15 items	10.58	3.33	.71	9.38	3.06	.62	11.04	2.05	.74	9.31	2.33	.62
20 items	13.78	4.15	.69	13.32	4.15	.67	15.41	2.48	.77	12.53	2.96	.63
25 items	16.98	5.18	.68	17.60	5.12	.71	19.84	3.09	.79	15.90	3.47	.64
30 items	20.86	5.91	.70	21.32	6.26	.71	24.16	3.80	.81	18.80	4.51	.63
Interview Criterion measure	**Mean**	**SD**	**PC**	**Mean**	**SD**	**PC**	**Mean**	**SD**	**PC**	**Mean**	**SD**	**PC**
100 items	62.20	22.04	.62	55.53	24.23	.56	65.14	16.21	.65	43.76	17.44	.44

PC = proportion correct mean score

Then the retrospective interviews were analysed trying to get closer insights into how participants arrived at the correct answers in the MCT. Finally, we looked at how the scores on the MCT relate to the overall score of the interview measure.

In the first analysis, participants' scores on the original 10 MCT items were compared with their knowledge of the same 10 words in the interview. The result

3. An anonymous reviewer wondered whether the proportion correct scores in Table 3 could be interpreted in terms of overestimation of the 10 MCT items vs. the 30 MCT items. While there is some difference, especially for the 6K B administration, the test versions with the additional items added are generally very close to the original items in proportion correct scores. More importantly, the 10 item vs. 11–30 item discussion that comes later in the paper is not about overestimation, but rather about sampling rate. Our discussion about overestimation all derives from data comparing the original 10 MCT items with the criterion interview measure of the same 10 items and an additional 90 items, respectively. These are reported in Tables 4–6, and Figures 3a-3d.

of this analysis is a direct comparison between the ability to answer a MCT item correctly and demonstrated knowledge of the meaning recall of the corresponding target word, which should permit understanding that word in a reading context. In the contingency matrices (Table 4), desirable results are shown in cells A and D, where there is congruency between item performance and demonstrated knowledge. Across the four data samples, a range of 70–84% of the responses fell into these categories. However, the B and C cells are problematic, as they indicate discrepancies between the MCT and the criterion interview measure. Across the four data samples, there is a relatively small proportion of cases where the participants were unable to answer the MCT item correctly, although they demonstrated knowledge of the word in the interview (Cells B: 2, 3, 4 and 8%), hence underestimation does not seem to be a major issue for the MCT format or at least in case of the VST. In terms of overestimation, i.e. where participants were credited with knowing a target word on the MCT, but where they could not demonstrate adequate meaning knowledge in the interview, the values are considerably higher (Cells C: 11, 13, 18 and 26%). In absolute terms, these figures indicate that 11–26% of the items on the MCT are gained without the requisite knowledge about the target words.

We ran an analysis to see whether the above percentages of overestimation were higher than could be expected from simple blind guessing. First, we determined the number of test words which were actually unknown according to the interview (e.g. 7 words). By using blind guessing as a strategy, a participant could be expected to answer .25 of unknown words correctly, so we calculated a potential guessing figure (e.g. $7 \times .25 = 1.75$). We then compared this guessing figure (1.75) with the number of unknown words answered correctly on the MCT (Cell C cases) (e.g. 4). For each frequency band, we checked whether the average overestimation value (e.g. 4) was significantly higher than the average potential guessing score (e.g. 1.75) through paired-samples t-tests. The results showed that this was the case for two frequency bands: 3K ($M = 1.84$, $SD = 1.34$ vs. $M = 1.19$, $SD = .58$, $t(49) = 3.81$, $p < .001$) and 9K (($M = 2.59$, $SD = 1.58$ vs. $M = 1.41$, $SD = .48$), $t(50) = 6.23$, $p < .001$). For the 6K band, the results approached significance for one administration (6K-B) ($M = 1.06$, $SD = 1.16$ vs. $M = .84$, $SD = .47$), $t(50) = 1.82$, $p = .075$), while not achieving significance for the other administration (6K-A) ($M = 1.32$, $SD = 1.32$ vs. $M = 1.13$, $SD = .60$), $t(39) = 1.14$, $p = .261$). Overall, Table 4 shows that there is clear overestimation, and the above analysis shows that there is a tendency for this overestimation to be even greater than could be expected from blind guessing.

As blind guessing does not explain why test-takers answered a relatively high number of items correctly when also showing no meaning recall knowledge of the target words, it suggests that test-taking strategies were employed by the

Table 4. Contingency matrices of MCT scores and interview scores for the same 10 words from the four data sets.

		MCT 3K		MCT 6K (A)		MCT 6K (B)		MCT 9K	
		Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
Interview	**Knew**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**
		252 (50%)	10 (2%)	206 (52%)	13 (3%)	298 (58%)	40 (8%)	199 (39%)	23 (4%)
	Did not know	**C**	**D**	**C**	**D**	**C**	**D**	**C**	**D**
		92 (18%)	146 (29%)	53 (13%)	128 (32%)	54 (11%)	118 (23%)	132 (26%)	156 (31%)

The four table cells for each data set contain the absolute and relative numbers for the following type of responses across the MCT and Interview measures:
Cell A = Answered MCT item correctly and demonstrated knowledge in interview
Cell B = Answered MCT item incorrectly but demonstrated knowledge in interview
Cell C = Answered MCT item correctly but did not demonstrate knowledge in interview
Cell D = Answered MCT item incorrectly and did not demonstrate knowledge in interview

participants. To investigate this further, we analysed the strategies used by 20 test-takers while answering the 10 original MCT items in the Lithuanian 3K study (i.e. 200 cases). Based on Paul, Stallman and O'Rourke (1990) and Schmitt, Ng and Garras (2011), we identified 6 main test-taking strategies:

1. **Knowing the meaning:**[4] test takers stated that they knew what the word means.
2. Inferring the meaning, when a **member of a word family** is known (e.g. participants chose the correct meaning of the word *to pave*, because of knowing the word *pavement*)
3. **Elimination and association**: test takers eliminated the answers that they thought were illogical ("*given golden edges* seems strange"), or that they thought they knew the word for ("I thought of all the synonyms meaning travelling and *rove* is not one of them") or used various associations ("I thought *jug* sounds similar to *jar* and *mug*"). Test takers nearly always reported using elements of these two strategies together.
4. Inferring from **similar word forms in the test items**: this strategy was used for two items in particular: Item 6 (word *strap* associated with the word *strip*) and Item 10 (*lonely* associated with *lonesome*).
5. Inferring the meaning based on the **context of the sentence**.
6. **Blind guessing**.

There were also other strategies used, but they were very uncommon (only 9/200), for example, choosing the shortest option (1 time), choosing the option that is understood best (1 time), choosing the definition which sounds best (1 time). However, they never led to correct test answers. Table 5 summarizes how often and how successfully various strategies were used by the test takers. Knowing the meaning (Strategy 1) was the most frequent reason for choosing an answer, used by all 20 participants. It resulted in selecting the correct answer most of the time and showed meaning recall knowledge as indicated by the interview. The least frequent strategy was blind guessing (Strategy 6), used by only 5 participants. As in other studies (e.g. Rupp, Ferne, & Choi, 2006), it seems that blind guessing was usually seen as a last resort, when other test-taking strategies (in this case, trying to think about word associations or finding test-related clues) failed to produce a result. Hence, as blind guessing seems to be infrequent and unsuccessful, it does not appear to distort the test scores, at least based on this limited data sample.

The strategy of inferring word meaning from other members of the word family (Strategy 2) seems to be very successful and it demonstrates actual knowledge

4. Strictly speaking, knowing a word is not a test-taking strategy, but we refer to it as such for ease and conciseness of discussion.

most of the time. This is a positive result, which shows that the notion of the word family does work in the VST, as the knowledge of one member of the word family can lead to answering the test item correctly.

The most widely used test-taking strategy was elimination and association (Strategy 3). This combined strategy included both unclear intuitions and partial knowledge, and was much more successful than simply blind guessing. Of 53 cases, learners were able to answer the MCT items correctly in 20, despite only knowing the word in 6 cases. Thus, by using this strategy, test-takers were able to answer 14 out of 53 items correctly (26%), while demonstrating no knowledge on the interview measure. The use of this strategy was also widespread, as most individual test takers managed to answer one or two items correctly by employing it.

Strategies 4 and 5 involve the use of test clues. Strategy 5 is not detrimental, because guesses based on the context of items were extremely unsuccessful, which shows that the contexts are minimal and that all of the options are plausible with those contexts. Strategy 4, on the other hand, is more problematic, especially for this set of items. It was mostly used for two specific test items and in both cases words similar to the target words were included in correct options (*strip — strap*: orthographic similarity, *lonesome — lonely*: same stem), which made this strategy extremely successful. For the word *lonely*, where the connection between the target word and the definition word was meaning based, most of the time the examinees reasoned this connection out and answered the item correctly. While some of the participants may have had some partial knowledge of this item, some others might have simply inferred the connection from the formal similarity of the two words (*lonesome — lonely*). Therefore items that have similar distractors should be avoided, because the test is supposed to check learners' existing knowledge of the words rather than evaluating their inferencing abilities (or providing learning opportunities from this inferencing). As for the word *strap*, participants who answered the item correctly based on this strategy showed no knowledge of the word. While it could be argued that the *lonely* item might capture some type of lexical knowledge in some cases, the *strap* item does not. However, Strategy 4 could probably be eliminated as a viable strategy with careful analysis and revision of the MCT test items.

Summarizing, the analysis of test-taking strategies seems to suggest that the investigated MCT is not very affected by blind guessing, as test-takers do not seem to guess blindly that often. Rather the test seems to be prone to various other test-taking strategies, where at least some of these must be seen as construct-irrelevant.

While the previous two analyses looked only at the original VST items and if the correct answer in the test showed that participants could recall the meaning of the target word or simply relied on test-taking strategies to arrive to the correct answer, the final analysis looks at how the test-score of the original VST matches

Table 5. Test-taking strategies and success rate.

	Test-taking strategies					
	1. Knowing the word	**2. Knowing family member**	**3. Elimination and association**	**4. Similar words**	**5. Context sentence**	**6. Blind guessing**
Number of participants using strategy[a]	20 (100%)	8 (40%)	17 (85%)	13 (65%)	9 (45%)	5 (25%)
Frequency of use[b]	95	8	53	18	10	7
Getting correct answer on the MCT	91 (96%)	7 (88%)	20 (38%)	18 (100%)	1 (10%)	1 (14%)
Knowing the word on the interview measure	90 (95%)	6 (75%)	6 (11%)	11 (61%)	0 (0%)	0 (0%)

[a] Out of 20 participants
[b] Out of 191 answers

the knowledge of the relevant frequency band. For this purpose, the participants' scores on the original 10 MCT items were compared with their performance on the 100-word interview for the four data sets. In Figures 3a-d, the plots show these comparisons. As 100 words for the interview are sampled from the same population as the 10 items for the original MCT only with a different sampling ratio (1:10 for the interview, compared to 1:100 for the MCT), in the ideal case the test score from the original 10 items on the MCT (x-axis) should predict the vocabulary interview measure score (y-axis) as closely as possible. However, from these figures we see that that there is often a considerable mismatch between MCT scores and the interview measure scores. The plots illustrate that there is both under- and overestimation, but that the latter is clearly dominant, especially for the 3K, the mixed-L1 6K, and the 9K. In order to go beyond mere impressionistic observation based on these plots, however, we carried out paired sample t-tests based on mean proportion correct scores for the MCT and the Interview. Table 6 shows the results.

As can be seen, for 3 out of 4 data sets, the MCT proportion correct scores are significantly higher than the Interview proportion correct scores, with large (or approaching large) effect sizes. The fourth data set (6K (B)) showed the same trend, although the p-value fell just short of .05.

RQ2. Is the sampling rate of the MCT sufficient to represent test-takers' knowledge of a relevant frequency band in a valid way? If not, what would a more appropriate sampling rate be?

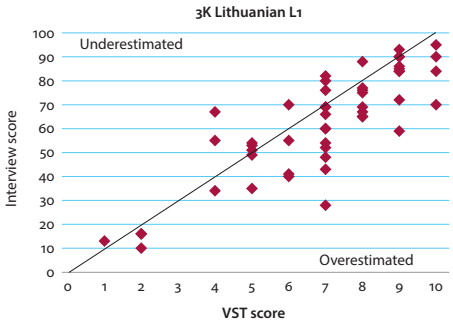

Figure 3a. Comparison of test scores on 10 MCT words (X-axis) and interview scores for 100 words (Y-axis) for the 3K frequency band.

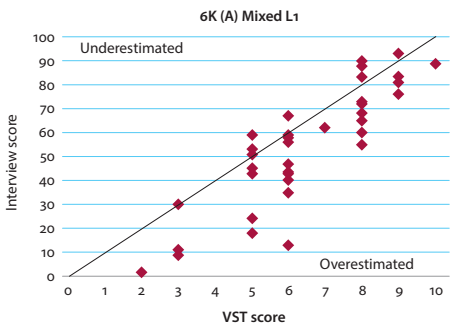

Figure 3b. Comparison of test scores on 10 MCT words (X-axis) and interview scores for 100 words (Y-axis) for the 6K frequency band.

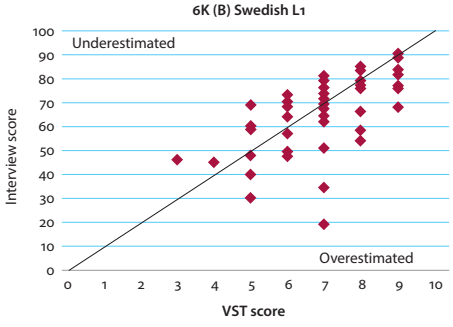

Figure 3c. Comparison of test scores on 10 MCT words (X-axis) and interview scores for 100 words (Y-axis) for the 6K frequency band.

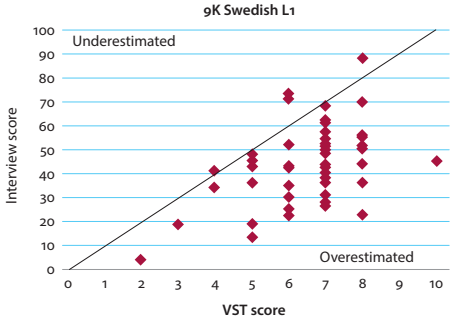

Figure 3d. Comparison of test scores on 10 MCT words (X-axis) and interview scores for 100 words (Y-axis) for the 9K frequency band.

The *Vocabulary Size Test Instructions and Description* document specifies that examinees' scores should be multiplied by 100 to find their total vocabulary size, based on the test design principle that one item on the test represents 100 items in the relevant frequency band. This design principle entails that items from a particular level represent words from the frequency band from which they were sampled, and not from other frequency bands (e.g. a 3,000-level item cannot be extrapolated to indicate vocabulary size at the 6,000 frequency level). The question then becomes how many test items are necessary to represent each distinct 1,000-word frequency band. The results reported in the previous section seem to suggest that a sample of 10 items for a 1,000-word frequency band seems to overestimate test-taker's knowledge.

In order to address this question further, a correlation analysis was carried out. The participants' scores on the original 10 MCT items were correlated with

Table 6. Pair-wise comparisons (t-test) of proportion correct means for MCT scores and Interview scores.

Data sets	Proportion scores		t-test statistics			
	M	SD	df	t-value	P	Effect size
3K MCT 10 words	.688	.223	49	3.73	.001	.47
3K Interview 100 words	.622	.220				
6K (A) MCT 10 words	.648	.188	39	4.72	.000	.60
6K (A) Interview 100 words	.555	.242				
6K (B) MCT 10 words	.690	.140	50	1.98	.053	.27
6K (B) Interview 100 words	.651	.162				
9K MCT 10 words	.649	.158	50	9.51	.000	.80
9K Interview 100 words	.438	.174				

their scores on the 100-word interview for the three frequency bands (Table 7). A Spearman's rho correlation was used as a Kolmogorov-Smirnov test indicated that some score distributions deviated from normality. The correlations ranged from .50-.86. While these correlations might seem acceptable in other language testing contexts, the fact that a single test item represents 100 words means that very high correlations are required to avoid large error in the resulting vocabulary size estimate. With R_s^2 values ranging from .25–.74, ten items do not appear to provide the desired level of precision.

In order to ascertain whether additional test items would improve the level of precision, five items at a time were added in an incremental fashion, and correlation values obtained for 15, 20, 25 and 30 items. This resulted in fairly consistent

Table 7. Correlations of the MCT item scores and Interview measure scores.

	Interview measure scores							
	3K Test administration		6K Test administration A		6K Test administration B		9K Test administration	
	r_s	R_s^2	r_s	R_s^2	r_s	R_s^2	r_s	R_s^2
MCT								
10 items	.81**	.66	.86**	.74	.66**	.44	.50**	.25
15 items	.85**	.73	.94**	.88	.80**	.64	.69**	.48
20 items	.86**	.74	.89**	.79	.86**	.74	.66**	.44
25 items	.88**	.77	.88**	.77	.85**	.73	.79**	.62
30 items	.91**	.83	.95**	.90	.89**	.79	.85**	.73

** All correlations Spearman's rho, significant at $p < .01$

improvements in correlation values as additional items were added. In absolute terms, 30 items generated R_s^2 values ranging between .73-.90. This evidence demonstrates that adding additional items to each test level results in better estimates of the 1,000-word bands.

Discussion

With MCTs being widely used for measuring vocabulary size, but with existing validity evidence relying mainly on computer simulations and probabilistic modelling, the present study was designed to investigate two central issues concomitant with vocabulary MCTs: the potential problem of overestimation (and underestimation) and that of sampling rate. To do this, we designed a case study featuring a fairly recent but influential four-option MCT: the VST. Using a test score interpretation approach, our criterion-related validity study compared test-takers' scores on the MCT test with their performance on another test instrument targeting the same knowledge. To the best of our knowledge, no studies to date have used this approach when researching this particular MCT. Two research questions guided our investigation.

Our first research question explored the degree to which answers on the MCT items reflect demonstrable knowledge of the target words' meanings. The results show that there was a clear tendency for scores on the MCT to be proportionally higher than scores from the interview measure. There seem to be two principal ways in which these results can be interpreted. The first interpretation would hold that the difference between test-takers' scores on the MCT and their scores in the oral interview comes from the fact that the task in the MCT is a meaning recognition task, whereas the task in the interview is a meaning recall task, where the latter is typically a more demanding task. As a case in point, Laufer & Goldstein (2004) found that meaning recall (called "active recall" in their study) was a more difficult task than meaning recognition (called "active recognition" in their study) when testing the vocabulary knowledge of 435 learners of English. The first possible explanation for the obtained results would therefore be that the discrepancy in scores stems simply from testing different aspects of word knowledge in the two measures.

However, ascribing the difference to only a dissimilarity in tasks and stopping there is too simplistic in our view, and disregards the score interpretation rationale employed in the study. This approach highlights the fact that practitioners and researchers want to use the score to be able to say something meaningful about a test-taker's ability. No end-user will want to interpret scores from a MCT as the ability to choose among alternatives on M-C items. Rather, teachers, researchers,

and administrators will make inferences regarding language use based on the test scores. For example, they might want to know if a person has the vocabulary necessary for reading unsimplified authentic texts. The MCT taps into meaning recognition, whereas understanding a text during reading requires meaning recall. In the reading situation, no meaning alternatives are provided — word forms in the text must trigger the activation of corresponding meanings. We do not think it is farfetched to assume that a test-taker with an estimated vocabulary size of e.g. 8,000 word families will be expected to be able to read texts like novels and newspapers with an adequate comprehension level, just as predicted in Nation (2006). However, if the extrapolated score of 8,000 words is based on an overestimation, and in the worst case a quite substantial one, then our hypothetical test-taker may experience difficulties when reading authentic texts, counter to predictions.

On balance, then, our results may provide more or less cause for concern depending on which of the two main interpretations is given priority, and in effect how scores from a vocabulary size MCT like the VST are interpreted and used. On a general note, we think that the use of MCTs as measures of vocabulary size would benefit from being subjected to an argument-based approach to validation (Kane, 2013), whereby interpretation and use of test scores are made explicit, followed by an evaluation of the plausibility of these proposals.

In line with the test score interpretation and use argument from above, if we look at the obtained results, in 11–26% of all cases concerning the original items from the investigated MCT participants were credited with meaning knowledge of these words that they did not possess. When comparing scores on the ten original items with knowledge of the 100 words on the interview measure, the same pattern of discrepancy obtains. In his test specification document, Nation (2012) does acknowledge that a score on the MCT under investigation here, the VST, is a "slightly generous estimate of vocabulary size". In our view, our research shows that overestimation, for this is what he means, is more problematic than Nation suggests. While it is true that the VST can provide a rough estimate of vocabulary size, it is debatable whether it has the level of precision necessary for many pedagogical purposes, let alone research purposes where a very high level of accuracy is needed. Perhaps our interpretation of overestimation should not come as a surprise, though, given the problems that previous researchers have identified with various multiple-choice vocabulary tests. In the research reviewed in the background section, we saw that a six-choice format like the Vocabulary Levels Test (VLT) is likely to lead to an average score increase of 16.7% due to guessing for most levels of ability (Stewart and White, 2011). Stewart (2014) concluded that the VST format is fraught with similar overestimation tendencies, namely inflation levels as high as 25% for most ability levels. This number, computed through a formula based on elementary probability theory, lies interestingly close to the

percentage of overestimation for the 9K band observed in the present study with real examinees (26%). Although our study is a case study of three of the sections of the test, and our position to generalize somewhat limited, our results still point to tangible issues that at least call for more follow-up studies.

Our analysis is a post-hoc investigation, which was constrained by the test design principles of the extant MCT version, namely that the 10 words sampled for each test section represent the 1,000 words in the relevant frequency band, which in turn entails each target word representing 100 words from the lexical population. This required a validation approach which compared performance on target items on the MCT with knowledge of the words in the relevant 1,000-word frequency band. However, in future development of new vocabulary size tests, or revision of existing tests, it is interesting to think whether the notion of 'difficulty' could be profitably exploited, as opposed to frequency being the sole criterion for target word selection. While frequency will always be relevant to the design of vocabulary size tests, it can only be an indicative approximation of actual word occurrences in a language. We therefore wonder whether word difficulty (e.g. as indicated by Beglar (2010) using IRT modeling, and Gyllstad (2012) using classical item analysis) can be used as a supplement to frequency in the process of ranking and sampling target words for inclusion in future tests. However, the test in our case study was not constructed on this premise and arguably then cannot be evaluated along these lines. Ultimately, the inclusion of difficulty is an empirical question, and needs to be evaluated for potential benefits and limitations.

In our study, we also investigated the presence and effect of test-taking strategies. We did this through a retrospection interview with 20 of the participants in the 3K section sub-study. Unsurprisingly, our participants actively used a range of test-taking strategies when attempting to answer items that they did not know. Overall, the retrospective interviews showed that blind guessing was both uncommon and unsuccessful, and so has little effect on the resulting MCT scores. This is in a way good news, as it can be used as an argument against introducing some kind of correction formula for blind guessing (see e.g. Eyckmans, 2004; Huibregtse, Admiraal & Meara, 2002). Another approach in a four-choice option is to introduce an "I don't know" option. Zhang (2013) has shown that such measure reduces the number of random guesses on the test, but also that it discourages attempts based on partial knowledge. Paul Nation (personal communication) claims that guessing on vocabulary tests is even desirable, because it is likely to draw on sub-conscious knowledge, and potentially shows the partial knowledge of the target words. We think that depends on what is meant by 'guessing'. In the retrospective interviews, we saw that the use of knowledge of the word family of the target word typically helped examinees answer items correctly. This test-taking strategy probably does draw on partial knowledge, which is useful and can

be applied in language communication. But when the strategy of elimination and associations is used, and in our case it frequently was, it is actually rather difficult to make any claims about partial knowledge. This strategy must be seen as an undesirable, construct-irrelevant strategy that is problematic from a validity point of view. One remedy for the widespread use of elimination strategies is to pay more careful attention to distractor construction, and to use distractor analysis as part of the development/validation process. A further option is simply increase the number of options in the MCT. Stewart (2012) observed that scores dropped considerably (close to 40%) and reliability increased when introducing as many as 25 options in the first three 1,000 word test sections of a VST format. The drawback, of course, is that such a measure heavily restricts the number of frequency bands that can be tested in one administration.

Our second research question asked whether the sampling rate of 1:100 of the investigated vocabulary size MCT is sufficient to adequately represent knowledge of words in the relevant 1,000-word frequency band. Our results (see Tables 4 and 6, and Figure 3) show that the ten items per test section are able to provide a rough indication of vocabulary size that may be sufficient for test purposes that only require a ballpark figure, and where overestimation is not problematic (e.g., a placement purpose which divides learners into a higher and a lower group). However, many purposes require more precision. For example, the pedagogical purpose of testing students' vocabulary size in order to select the appropriate graded reader level requires vocabulary size estimates accurate to within around 500 word families.[5] Furthermore, in incidental vocabulary acquisition research from reading, gains are typically quite small (e.g. 4%, Waring and Takaki, 2003), which necessitates accuracy in measuring very small numbers of gain words in order to demonstrate this learning. Our investigation into the behaviour of one MCT at three frequency bands strongly suggests that this test is likely to produce a degree of measurement error (e.g. 16%-30%) which is far greater than typical incidental gains (e.g. the abovementioned 4%), on the assumption that the item results are interpreted at the meaning recall level of mastery. Many studies of incidental vocabulary learning have used 4-option multiple-choice items to measure learning of target words (e.g. Horst, Cobb, & Meara, 1998; Saragi, Nation, & Meister, 1978), and our study indicates that these items may not have the level of precision required for such research purposes.

However, our results showed that an increase in the number of items in each test section can lead to a more accurate representation of the corresponding frequency band (as evidenced in Table 7). We found that having 30 items per

5. For example, the *Cambridge Discovery Readers* series has the following levels: Starter: 250 headwords, 1: 400, 2: 800, 3: 1,300, 4: 1,900, 5: 2,800, 6: 3,800.

frequency band led to r^2 values of .73–.90, which may well be adequate for many uses, although this would have to be established for each individual test purpose. This is of course in line with CTT assumptions, but we must remember that this only holds if the lengthening of the test is done through adding items of sufficient quality. Adding more items generally did improve the test performance, and 30 items in total produced correlations as high as .85-.95.[6] We do not know whether even more items would add to the test's accuracy, although we suspect that more than 30 items would entail practicality issues, given that the VST contains 14 separate levels.

We used the VST as a case study to explore the characteristics of multiple-choice vocabulary size tests. Our results indicate cause for concern (if one accepts the score interpretation line of argument), but it is important to note that these concerns probably apply to M-C formats as a class, especially those with low item sample size, rather than being limited to the particular MCT investigated here. At a minimum, we feel our results indicate that M-C formats in general need to be more carefully scrutinized in terms of their ability to provide the kind of vocabulary information practitioners and researchers require.

Conclusion

Vocabulary size MCTs have a general appeal and are widely used for a number of different assessment purposes. However, they are not free from problems. Results from our case study show that there was a mismatch between MCT scores and scores on a carefully designed criterion measure. We proposed two main interpretations of these results, one which ascribes the mismatch to a difference in tasks used in the two measures, and one which argues that a test score interpretation and use perspective is essential, from which follows that the results must be seen as a tendency of overestimation of MCT scores in relation to demonstrable word knowledge. Our case study results also show that a higher sampling rate than 1:100 is needed in order to better represent the underlying population of words in the corresponding frequency bands. Finally, our methodology illustrates that external validity evidence (e.g. interviews and retrospective protocols) can provide a useful, and perhaps necessary, complement to other validation approaches in order to get a more comprehensive picture of the characteristics of vocabulary size MCTs.

6. It should be noted that there was a greater overlap of items (words) between the 30-item MCT version and the 100-word interview than between the 10-item MCT version and the interview, something that may at least partially have affected the correlation coefficients.

References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

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

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, 27(1), 101–118. DOI: 10.1177/0265532209340194

Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, 61(2), 367–413. DOI: 10.1111/j.1467-9922.2010.00613.x

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

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, 8(4), 341–349. DOI: 10.1037/1040-3590.8.4.341

Eyckmans, J. (2004). *Measuring receptive vocabulary size: Reliability and validity of the yes/no vocabulary test for French-speaking learners of Dutch*. Utrecht: LOT.

Gyllstad, H. (2012). Validating the Vocabulary Size Test. A classical test theory approach. Poster presented at The Ninth Annual Conference of EALTA, Innsbruck, Austria, 31 May – 3 June. Retrieved from: <http://www.ealta.eu.org/conference/2012/posters/Gyllstad.pdf>

Horst, M., Cobb, T., & Meara, P. (1998). Beyond A Clockwork Orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language*, 11, 207–223.

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

Kamimoto, T. (2008). *Nation's vocabulary levels test and its successors: A re-appraisal*. Unpublished PhD Thesis. University of Wales, Swansea.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. DOI: 10.1111/jedm.12000

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

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

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams (Eds.), *Performance and competence in second language acquisition* (pp. 35–53). Cambridge: Cambridge University Press.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, 4(2), 142–154. DOI: 10.1177/026553228700400202

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749. DOI: 10.1037/0003-066X.50.9.741

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes*, 63(1), 59–82. DOI: 10.3138/cmlr.63.1.59

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Heinle & Heinle.

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

Nation, I. S. P., & Gu, P. Y. (2007). *Focus on vocabulary*. Sydney: NCELTR Publications.

Nation, P. (1993). Using dictionaries to estimate vocabulary size: essential, but rarely followed, procedures. *Language Testing*, 10(1), 27–40. DOI: 10.1177/026553229301000102

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

Paul, P. V., Stallman, A. C., & O'Rourke, J. P. (1990). *Using three test formats to assess good and poor reader's word knowledge*. Technical Report No. 509 of the Center for the Study of Reading. University of Illinois. Retrieved from <https://www.ideals.illinois.edu/bitstream/handle/2142/17704/ctrstreadtechrepv01990i00509_opt.pdf?sequence=1>

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

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. J. Kunnan (Ed.) *Validation in language assessment* (pp.41–61). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, 23(4), 441–474. DOI: 10.1191/0265532206lt337oa

Saragi, T., Nation, I. S. P., & Meister, F. (1978). Vocabulary learning and reading. *System*, 6, 72–78. DOI: 10.1016/0346-251X(78)90027-1

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Houndmills: Palgrave Macmillan. DOI: 10.1057/9780230293977

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition*, 19, 17–36. DOI: 10.1017/S0272263197001022

Schmitt, N., Ng, J. W. C., & Garras, J. (2011). The word associates format: Validation evidence. *Language Testing*, 28(1), 105–126. DOI: 10.1177/0265532210373605

Schmitt, N., & Schmitt, D. (2012). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, available on CJO2012*. DOI: 10.1017/S0261444812000018

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. DOI: 10.1177/026553220101800103

Stenius Staehr, L. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, 31, 577–607. DOI: 10.1017/S0272263109990039

Stewart, J. (2012). A multiple-choice test of active vocabulary knowledge. *Vocabulary Learning and Instruction*, 1(1), 53–59. DOI: 10.7820/vli.v01.1.stewart

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

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

Uden, J., Schmitt, D., & Schmitt, N. (2014). Jumping from the highest graded readers to ungraded novels: Four case studies. *Reading in a Foreign Language*, 26(1), 1–28.

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader. *Reading in a Foreign Language*, 15(2), 130–163.

Weir, C. J. (2005). *Language testing and validation*. Houndmills: Palgrave McMillan.

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review*, 53(1), 13–40.
Wood, B. (1928). *New York experiments with new-type modern language tests*. New York, NY: Macmillan.
Zhang, X. (2013). The I don't know option in the Vocabulary Size Test. *TESOL Quarterly*, 47(4), 790–811. DOI: 10.1002/tesq.98

Author's addresses

Henrik Gyllstad
Lund University
Centre for Languages and Literature
Box 201
221 00 Lund
Sweden

henrik.gyllstad@englund.lu.se

Laura Vilkaitė
University of Nottingham
School of English
Trent Building
University Park
Nottingham
NG7 2RD
UK

laura.vilkaite@nottingham.ac.uk

Norbert Schmitt
University of Nottingham
School of English
Trent Building
University Park
Nottingham
NG7 2RD
UK

norbert.schmitt@nottingham.ac.uk