The importance of selecting appropriate instruments when measuring receptive vocabulary size

Irene Castellano-Risco
Universidad de Extremadura
ircastellano@unex.es

Abstract: Recent trends in language teaching point to the importance of vocabulary in language proficiency (Boers & Lindstromberg, 2008; Milton, 2009; Schmitt, 2008; Nation, 2010). Due to its importance, research in the field of vocabulary acquisition has yielded different instruments that can be used for analysing receptive vocabulary size (Nation & Beglar, 2007; Meara, 1992; Schmitt, Schmitt & Clapham, 2001). This study explores different receptive vocabulary tests with the aim of exploring the impact of using different tests on the results. Two instruments, the checklist tests (Meara, 1992) and the Vocabulary Levels Test (VLT; Schmitt, Schmitt & Clapham, 2001), were used in this study. A total of 138 students in their third year of secondary education were asked to complete the first level of both tests, and the data were scored following different formulae. In the case of the VLT, the formula proposed by Schmitt, Schmitt, and Clapham (2001) was used. For the checklist tests, two formulae were employed: the one proposed by Meara and Buxton (1987) and one developed by Meara (2010). Finally, those results were compared with each other and with other studies. In light of the results, the selection of the instruments has a direct impact on the results obtained.

Keywords: VLT, checklist tests, receptive vocabulary size, vocabulary measurement.

La importancia de seleccionar instrumentos apropiados para medir el tamaño de vocabulario receptivo

Resumen: En las últimas décadas, el vocabulario se ha erigido como parte esencial en el desarrollo de una lengua para un correcto desarrollo de la misma (Boers & Lindstromberg, 2008; Milton, 2009; Schmitt, 2008; Nation, 2010). Debido a esta importancia, la investigación en la adquisición de vocabulario ha desarrollado una serie de instrumentos que pueden ser utilizados para medir el tamaño de vocabulario receptivo (Nation & Beglar, 2007; Meara, 1992; Schmitt, Schmitt & Clapham, 2001). Este estudio analiza diversas herramientas con el objetivo de observar si hay diferencias en las mediciones de vocabulario, y si las hubiera, para seleccionar cuál de ellas sería más adecuada para utilizar con alumnos de Educación Secundaria. Específicamente, dos tests, el Checklist (Meara, 1992) y el Vocabulary Levels Tests (VLT) (Schmitt, Schmitt & Clapham, 2001) fueron usados en este estudio. Se les administraron el primer nivel de ambos cuestionarios a un total de 138 alumnos cursando tercero de Educación Secundaria. Los resultados se obtuvieron haciendo uso de diferentes fórmulas: en el caso del VLT, se utilizó la fórmula propuesta por Schmitt, Schmitt y Clapham (2001). En el caso del checklist, se siguieron dos fórmulas diferentes: una propuesta por Meara y Buxton (1987) y otra desarrollada por Meara (2010). Finalmente, los resultados surgidos de la implementación de las diferentes fórmulas fueron comparados entre ellos y con otros estudios. En base a los resultados obtenidos se puede afirmar que la selección del instrument de medida tiene un impacto directo en los resultados obtenidos.

Palabras claves: VLT, checklist test, vocabulario receptivo, medición de vocabulario.

Recibido el 24/01/2018
Aceptado el 17/05/2018

1. Introduction

Nowadays, vocabulary knowledge is seen as one of the major factors that affects language learning (Boers & Lindstromberg, 2008; Schmitt, 2008; Milton, 2009; Nation, 2001). This vision of the teaching practice is fairly new. Until the late 1970s, vocabulary was relegated to a second place, studies on vocabulary were scarce (O'Dell, 1997, p. 258) and the few existing focused on methodological aspects of vocabulary instruction, rather than on vocabulary itself (Laufer, 1990). This tendency towards ignoring vocabulary teaching was directly applied to the teaching practice. For instance, vocabulary was not mentioned in syllabi or curriculums and teaching training materials often omitted vocabulary teaching and books (e.g. Lightbown & Spada, 1999). But, in the 1980s, studies on vocabulary became more and more common (Laufer, 1986, 1990; Meara, 1980, 1996a, 1996b; Nation, 1974, 1975, 1983, 1990; Richards, 1976; Xue & Nation, 1984), showing that the acquisition of vocabulary was central to master a language. This shift in the conception of language learning contributed to the blossoming of a new area of research, in which vocabulary constituted the central focus (Boers & Lindstromberg, 2008). Since that moment, and up to now, studies of vocabulary acquisition flooded, proving the key role that vocabulary plays in foreign language learning.

Based on its relevance, a large body of research has attempted to develop a number of instruments to measure vocabulary. A number of tests have emerged whose *raison d'être* is the analysis of vocabulary size that learners have and thereby the vocabulary learning process itself. However, those vocabulary tests present many variations that can lead to different results when examining vocabulary size. It is in consonance with this idea that this study, after a theoretical revision of the literature about this issue, examines the results yielded by different tests. The main objective is to explore how the use of different tests may produce dissimilar results and how these discrepancies in the results may affect vocabulary research itself.

2. Measuring receptive vocabulary size

The vocabulary construct is a more complex idea than it appears at first glance. In layman's terms, it can be defined as the number of words that comprise a language. However, a large number of studies have established that knowing a word involves many other aspects, such as knowledge of the word form, its structure, and its syntactic behaviour. To illustrate, Laufer (1990) has suggested that knowing a word includes learning the word form, word structure, syntactic

behaviour, meaning, and associative relations with other words. This view is consistent with Taylor's proposal, cited in López Campillo (1995), which defines word knowledge as knowledge of frequency of occurrence, style, register, dialect, semantic and syntactic collocations, morphology, semantics, polysemy, and translations. Similarly, Coady (1993) has concluded that knowing a word concerns knowing its syntactic behaviour, derivations, network of associations, semantic features, and the register in which it can be found. For Ur (1996), knowing a word involves the identification of the word form, grammar, collocations, word formation, and aspects of meaning.

Finally, the most comprehensible approach up to that moment was proposed by Nation (2005). He examined the previous proposals and classified most of the aspects previously suggested in three main categories of word knowledge: form, meaning and used. Each of these categories included more specific aspects. In the case of the category form, he identified not only the written and spoken forms but also the word parts that make up the words. As for meaning, he distinguished also between three aspects: the word and meaning, the references that the concept did and finally the different associations that learners may have when facing with the word. Finally, in relation to the dimension use, he identified as main aspects the understanding of the grammatical functions, the knowledge of collocations and the constrains on use. After this first categorization, he distinguished two dimensions of vocabulary in each of the aspects already mentioned: the receptive and productive form. By and large, Nation's proposal seems to include most of the agreements on the aspects involving vocabulary knowledge and is now taken as a framework of study.

Under the evidence that the dimensions of word knowledge correlate (Milton & Fitzpatrick, 2014), research on lexical competence measurement has focused on measuring one specific dimension, usually vocabulary size, rather than the whole of them. For that reason, when examining vocabulary tests, they are usually classified taking into account the vocabulary knowledge aspects that are measured. To illustrate, a well-known distinction in vocabulary studies is receptive vocabulary—the number of words that can be understood—versus productive vocabulary—the number of words that can be expressed (López, 1995). Another classification distinguishes between form-recognition tests—in which test-takers are asked to mark if they recognize a word— and meaning-recognition tests— that imply that the test-takers need to know, not only the form of the word, but the concept that is represented with the word.

Therefore, when examining the vocabulary tests, it is first necessary to examine whether the tests explore receptive or productive vocabulary. After that,

which vocabulary aspect is explored, should also be considered. Finally, other issues may be considered, such as the form, the length, or the aim of the tests.

In terms of receptive vocabulary size tests, one of the oldest methods is the checklist tests, also known as checklist tests due to the format used. These tests, developed by Meara and his colleagues in 1992, are by far the simplest for test-takers (Meara, 1992, 2010). They consist of a set of five tests that measure language knowledge ranging from the 2K band up to the 10K band. Test-takers simply read a list of lexical items and indicate whether they recognise each item. These tests' reliability has frequently been questioned; considering that they depend heavily on learners' perceptions, there is the chance that test-takers overestimate or underestimate their vocabulary knowledge. This problem of unreliability has been somewhat controlled by adding plausible pseudowords.

In 2007, the Vocabulary Size Test (VST) emerged, providing a new perspective on the area of vocabulary recognition tests. Developed by Nation and Beglar (2007), it is a multiple-choice meaning recognition test that aims to produce an overall vocabulary knowledge profile. Its objective stands in contrast to other receptive vocabulary tests, which present a diagnosis.

Nevertheless, although the use of those tests has been widespread, the one that excels above all the receptive vocabulary tests is the Vocabulary Levels Test (VLT; Schmitt, 2010). Developed by Schmitt, Schmitt, and Clapham (2001), it is a form-recognition matching test that focuses on vocabulary at five levels. Four of those levels are frequency levels that correspond to the number of word families considered sufficient to maintain a daily conversation (2,000 words), permit initial access to reading (3,000 words), enable independent reading (5,000 words), and permit advanced usage in most cases (10,000 words). The fifth level is focused exclusively on academic vocabulary (Schmitt, 2010).

2.1 Analysis of the tests

As seen in the previous section, there are several tests that can be used to measure receptive vocabulary size. However, subjecting the test-takers to all those different tests would be excessive, taking into account test-takers' attention span. This premise supports the need to select only two measuring tools. With that objective, the different instruments available for measuring receptive vocabulary size were analysed with respect to the following factors: bands of vocabulary measured, objective of the test, and time required to administer it.

Regarding the band of vocabulary measured in each test, some tools do not specify which band or what type of vocabulary is measured. Of those tests that do clarify this point, the VLT and checklist tests begin by measuring the 2K

band of vocabulary, while the VST also measures the 1K band. However, in the case of the checklist tests, information is only available about the band measured with the two first tests.

Although coverage of the 1,000 most frequent words is relevant, reviewing the results of other studies (Canga, 2013, 2015) with similar sample characteristics makes clear that secondary school students' overall results surpassed the 1K vocabulary band. Based on the assumption that secondary school learners have already acquired the 1K band, it is more meaningful to analyse the 2K band, so comparing the VLT and the checklist results would be a suitable choice as both of them present the same starting point. However, in the case of checklist tests —although it is explained that the first two levels measure the 2K band— it is not specified which band of vocabulary is measured in each level, so this disadvantage merits discussion.

As for the objectives of the tests, although all of them explore the receptive vocabulary size of test-takers, the VLT is not designed to estimate a person's overall vocabulary size. In fact, it is a diagnostic test, while the VST aims to measure overall vocabulary size. Finally, regarding the checklist tests, it cannot be assumed that receptive vocabulary knowledge is demonstrated.

Finally, another relevant factor for selecting the most appropriate tests is the time required to administer and score the test. If the final aim of the experiment —analysing the 2K band of receptive vocabulary— is taken into account, all the different tests must fit into an appropriate amount of time, so they cannot be individually time-consuming. However, if the objectives of the tests are borne in mind, the VST should be administered in its entirety, while for the rest of the tests, only the specific band to be measured may be administered. This fact makes the VST a relatively time-consuming test.

The strengths and drawbacks of all three tests, and therefore the reasons for choosing certain instruments over others, are summarized in the table below.

Table 1

	Bands	Objective	Advantages	Disadvantages
Checklist test	Up to 10K band	Diagnosis	- Easy and quick to take in class - Many items can be measured -Straightforward and automatic scoring	- No direct demonstration of knowledge. - Possibility of overestimation. -Bands not specified.
VLT	2K, 3K, 5K, 10K and Academic Band	Diagnosis	- Short definitions - Designed to tap into the initial stages of form–meaning link - Clusters designed to minimize aids to guessing - Academic vocabulary can be also measured.	- Not designed for providing an estimate of a person's overall vocabulary size.
VST	14 first bands	Measuring overall vocabulary size.	- It aims to measure overall vocabulary size.	- Too long.

As a result of the above considerations, the two tools chosen for analysing the receptive vocabulary size of the ninth-grade students were the VLT and checklist tests. TVLT provides information not only about whether learners recognise the words (form recognition) but also about meaning recognition. Moreover, it is not time-consuming, and academic vocabulary can be measured. The checklist test was also chosen because it measures a large number of items in a short period of time, although it has the disadvantage of uncertainty about which band of vocabulary is measured in each level.

As important as the selection of appropriate tools for measuring receptive vocabulary is knowledge of how those tests are scored. The following section focuses on the format and the scoring formulae used in each test.

2.1.1. Vocabulary Levels Test: features and scoring

The VLT is a form recognition test in which 30 items in 10 different clusters, each of which contains three definitions and six options, are presented in each level. Test-takers are asked to match the definitions with their corresponding words. The test is designed to minimize guessing aids, and all the words, those tested and those presented in the definitions, belong at most to the level being tested.

As for the scoring, a simple formula is used: the number of correct matches, or hits (H), is divided by the total number of definitions tested (TA; i.e., 30) and multiplied by 100. The result shows the percentage of coverage for this particular band of knowledge (Schmitt, 2010).

2.1.2. Checklist tests: features and scoring

The checklist tests are word recognition tests in which the test-takers have to mark known words with a Y and unknown words with an N (Meara, 2010). In order to avoid overestimation, the test includes false words, known as pseudowords. There are 60 words per level; 40 of these are real words, and 20 are pseudowords or non-words, (i.e., one out of three is a non-English word).

Focusing on the scoring of the test, the recognition of both real words and pseudowords should be examined. Four types of responses can be produced. Real words marked as known words are called hits (H). However, if pseudowords are marked as known, they are known as false alarms (FA). Real words unknown to test-takers are called misses (M). Finally, the pseudowords marked as unknown are called correct rejections (CR; Pellicer & Schmitt, 2012). Depending on the weight given to pseudowords, different formulae are used. This issue has been widely explored (e.g., Pellicer & Schmitt, 2012), and different formulae have been proposed since the advent of checklist tests. The first proposal consisted of subtracting the number of false alarms from the number of hits.

This formula was too simplistic. A solution came from the field of psychology with the development of signal detection theory, in which the aim is to quantify the ability to distinguish what is a real signal from noise, leading to a signal detection approach (Pellicer & Schmitt, 2012). This approach was used first by Anderson and Freebody (1983), and Meara and Buxton (1987) used the following set formula for scoring. This formula is known as the 'correction for

guessing' formula and is based on the proportion of hits to false alarms (Pellicer & Schmitt, 2012).

However, this formula overemphasised the hit rate over the false alarm rate. After a revision of his work, Meara (2010) proposed the use of a matrix he developed to convert hit and false alarm rates to a percentage-based vocabulary score. In this proposal, the number of false alarms is examined, and if either the number of false alarms is greater than 10 or the number of hits is below 10, the answer is invalidated.

In the following sections, both tests are implemented. However, not all the scoring formulae used in the checklist tests will be used, as some of them have already been rejected (Pellicer & Schmitt, 2012).

3. Research questions

The main purpose of this study is to analyse different instruments to explore whether the usage of different tests may affect the outcomes of research on vocabulary. With this objective, two research questions are posed:

> RQ1: Are there significant differences regarding whether the VLT or checklist test are used?
>
> RQ2: Are there significant differences regarding the use of different formulae when assessing the checklist tests?

4. Methodology

4.1 Context

This study was carried out in Extremadura, a monolingual region with a sparse population located in the south-western region of Spain on the border with Portugal. Those geographic features have influenced the way in which second language programmes have been implemented, resulting in, for example, the promotion of Portuguese learning and the implementation of content and language integrated learning (CLIL) programmes.

Participants came from four different urban secondary schools located in a medium-sized town (150,000 inhabitants). It was a convenience sample, made up of learners from the state schools that agreed to take part in the experiment. All the schools had different programmes to promote language teaching, including implementation of the CLIL approach and participation in European programmes—such as the Comenius and Erasmus programmes—with the

objective of using the foreign language with a communicative purpose as much as possible.

4.2 Participants

One hundred and thirty-eight students were included in this study. All the students were in their third year of secondary compulsory education (Year 9), with ages ranging from 14 to 16 years old.

Their EFL (English as a Foreign Language) learning background were heterogeneous as the type of instruction varied. On one hand, there were 72 students who had been exposed to the CLIL approach who had had more exposure to English (3,000 hours); on the other hand, there were 56 students who had attended only the EFL subject, so their exposure to English was more limited (1,200 hours).

4.3 Data collection and analysis

Tests were administered on two consecutive days: the 2K band VLT level and a version of the checklist tests were administered on this first day, and the academic vocabulary level of the VLT and another version of the checklist tests were given on the second day. Test administration was carried out in such a way as to avoid bias in the answers due to fatigue. Test-takers were asked to complete the tests and were informed that their information would be kept confidential. The instructions of the tests were given in both languages, English and Spanish, in order to ensure comprehension. Moreover, examples were also given.

For the VLT, test-takers selected the correct definition for the words given. As previously described, each level consisted of 30 words and 60 definitions. The time allowed to complete each test was seven minutes following instructions provided by the authors.

For the checklist test, test-takers had just five minutes to complete each test because, following Meara's instructions, this prevents students from taking 'too much time about individual items' (Meara, 2010, p. 13). Test-takers were given two different versions of the checklist tests following Meara's suggestion for collecting more reliable data. They were instructed to write a Y (yes) if they knew the word and an N (no) if they did not know the word or they were not sure.

All the results were analysed with SPSS V23 to check whether there were statistically significant differences between the results. The selected confidence interval was 95%.

5. Data analysis

5.1. RQ1: Are there significant differences regarding whether the VLT or checklist test are used?

For the results obtained from the checklist tests, two different instruments were used to score the tests: the formula proposed by Meara and Buxton (1987) and the matrix proposed by Meara in 2010. This enabled examination of how the use of different scoring instruments affected the final results.

When Meara and Buxton's formula (1987) was used, learners knew 80.89% of the 2,000 most frequent words, with a standard deviation (SD) of 17.66. These results were well above the 1K vocabulary band, as the estimated number of words known was 1,619 words.

However, when the matrix proposed by Meara (2010) was used, the results differed slightly. Learners knew 69.21% of the 2,000 most frequent words, with a SD of 22.5. In other words, learners knew approximately 1,384 out of the 2,000 most frequent words. These results were also above the 1K vocabulary band, but they were lower in comparison to the results obtained with the other formula. To analyse the significance of this difference, a U Mann-Whitney test was carried out. Results of the tests showed that there was a significant difference between both results ($p = .000$).

5.2 Results obtained from the VLT

According to the VLT, learners knew 57.49% of the 2,000 most frequent words. In relative terms, it could be said that learners knew 1,150 words. As can be seen in figure 1 below, this result was far below the results yielded by the checklist tests; however, third-year secondary school students' knowledge of the 2,000 most frequent words was still above the 1K band.

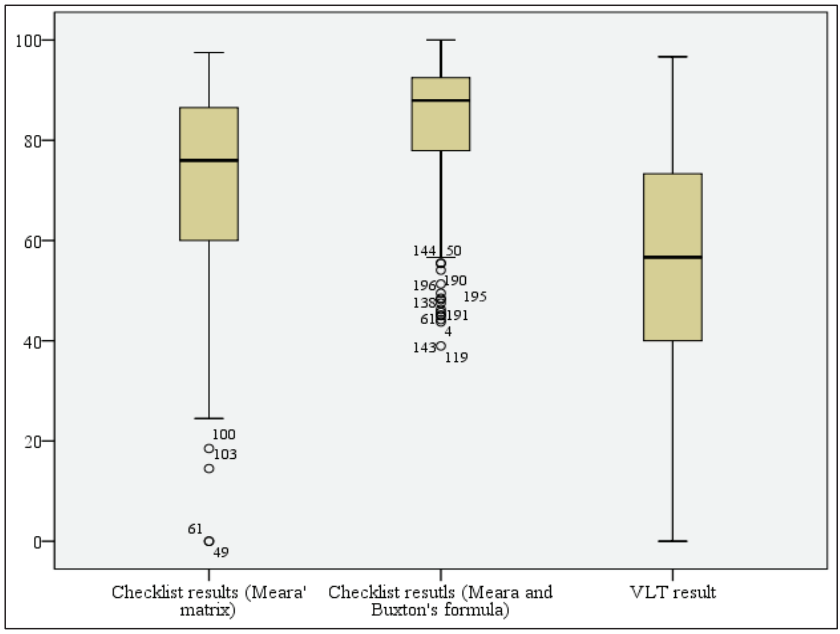

Figure 1: results of the VLT and checklist test

The difference in the results yielded by the checklist tests and VLT was also considered statistically significant (p =.000) according to the results of the U Mann-Whitney test, no matter the scoring formula followed. In other words, the choice of tests approach influenced the outcomes, as different approaches yielded results that varied significantly.

6. Discussion

This piece of work is an attempt to assess how the use of different tests by the same students may yield a wide range of vocabulary sizes and how the selection of an inappropriate tool may affect any study based on those results. Two different tools, the checklist tests and the VLT, were used to measure ninth-grade secondary school students, and their results were compared to look for differences regarding the test used.

Concerning the knowledge of the 2K band of receptive vocabulary, depending on the test used and the scoring method followed, different results arose. The results ranged from 57.49% in the VLT to 68.21% or 80.89% when using different score formulae proposed for the checklist test. All the differences—that

is, (1) differences between results from the VLT and the checklist tests and (2) differences arising from the use of divergent scoring formulae in the checklist tests—were statistically significant. From these results, it can be concluded that the tool selected seems to have some influence.

A number of reasons can explain these differences. Starting with the results of the checklist tests, the dissimilarities were expected. Meara (2010) realised that the original formula overestimated the hit rate, and that was the reason why he developed the new formula and the matrix. However, it was interesting to see to what extent this difference existed and whether it was related to the sample analysed. Moreover, it was also interesting to explore which of the options yielded outcomes more similar to those produced by other instruments, in this case, the VLT.

As for the variation between the VLT and the checklist tests, the difference in the results may be related to the aspects of vocabulary knowledge that measure each tool. The VLT is a meaning-recognition test, whilst checklist test is form-recognition test, so it is possible that learners recognised the form of the words measured in the VLT, but did not know the meaning.

Now that clear differences have been established, it can be identified which instrument is more in line with other studies in which secondary school students' receptive vocabulary size has been measured. Before starting, the first point that should be noted is that in Spain, studies analysing secondary school students' receptive vocabulary size are scarce. However, two studies (Canga, 2013, 2015) seem to have a sample that can be compared with the sample of this study. Participants of these studies were tenth-year learners (Canga, 2013, 2015), who had been learning English as a foreign language for ten years.

First, Canga (2013) used the VLT to analyse the receptive vocabulary size of 92 tenth-year EFL learners. These learners knew approximately 935 words out of the 2,000 most frequent ones. In other words, traditional EFL learners' receptive vocabulary size was below the 1,000 most frequent words.

These results also concurred with Canga (2015). He attempted to compare CLIL and EFL approaches, analysing tenth-year secondary school students and sixth-year CLIL learners. In order to examine the receptive vocabulary size, he also made use of the VLT. Focusing on the analysis of the EFL tenth-year secondary school students' coverage of the 2,000 most frequent words, he concluded that their receptive vocabulary size lay within the range of the 1,000 most frequent words, with an estimation of 936 words.

As previously pointed out, results from the present study were above the 1,000 most frequent words, regardless of the test chosen. However, when comparing

these results with Canga (2013, 2015), the results yielded by the checklist tests differed to a larger extent, whilst the VLT results were more similar, as they demonstrated that ninth-year secondary school students knew approximately 57% (1,140 words) of the 2,000 most frequent words.

Focusing on the comparison between Canga's samples and the results yielded from the VLT in this study, the results are similar perhaps due to the use of the same test in both studies. However, both samples differ in certain features, such as the hours of instruction in English and the kind of instruction.

The first difference between the two samples is the number of years learning English. Canga's sample (2015) was comprised of tenth-year secondary school students and the sample here analysed was made up of ninth-year secondary school students; the former should have been learning English for a longer period of time, resulting in a larger receptive vocabulary. However, this is not what is seen when comparing the outcomes, as the ninth-year learners obtained better results in comparison to tenth-year learners.

The second difference, and the one that may contribute most to the difference in the results, is the kind of instruction the test-takers were exposed to. In Canga's sample, all students were mainstream EFL learners; in the current sample, students were under the influence of different kinds of instruction. More than half of the test-takers were CLIL learners, and the rest were traditional EFL learners. The CLIL learners were exposed to more hours of instruction in English, which may have affected their receptive vocabulary size (Agustín, 2009; Jiménez & Ruiz de Zarobe, 2009; Ruiz de Zarobe, 2008; Canga, 2013; 2015). This can explain why the present sample had a larger vocabulary than Canga's sample.

The differences between Canga's sample and the checklist results were even greater. While Canga reported results below the 1K band, the checklist results surpassed the 1K band, with results between 70% and 80% of the 2K band depending on the scoring formula used. This difference may be related to the fact Canga made use of VLT to measure receptive vocabulary size.

Given these points and considering one of the main objectives of this study —examining how the use of different tests may affect the results of studies on secondary school learners' receptive vocabulary size—, it seems that the use of different tests and formulae definitely affects vocabulary research. In general, research includes comparison with other studies with similar samples. However, if the instruments used in the studies are not borne in mind when comparing the outcomes, the differences that may emerge cannot directly be attributed

to test-takers' vocabulary sizes, as these differences may also be related to the instruments selected.

7. Conclusion

This study has aimed to evaluate different receptive vocabulary size tests in order to (1) analyse whether the results of different instruments were consistent and (2) select the most suitable instrument as regards the sample analysed. After a pre-selection of the most appropriate tests with concern to time required, bands of vocabulary measured, and format, the focus was narrowed to two different tests: the VLT (Schmitt, Schmitt & Clapham, 2001) and checklist tests (Meara, 1992).

The experiment has highlighted the contrasting results of different receptive vocabulary size tests, so it can be concluded that the choice of a test affects the studies based on it. A more than 13% difference was found when contrasting the VLT results and the checklist test results obtained when using the Meara matrix (2010). This difference was even greater —nearly 25%— when comparing the VLT results with those obtained from the checklist tests scored with the formula proposed by Meara and Buxton (1987). Furthermore, differences between the checklist results were found when contrasting both ways of scoring, demonstrating the stress on the hit rate in Meara and Buxton's formula (1987).

In light of these findings, it could be stated that the kind of test chosen is likely to have an impact on the outcomes of the studies on vocabulary size. Different instruments may be measuring different aspects of word knowledge and it would definitely result in different findings. Therefore, it is essential to conduct a literature review in order to determine which instrument fits better with the aspects measured. Finally, the findings of this study are also relevant when comparing results of different studies.

References

Agustín, M.P. (2009). «The Role of Spanish L1 in the Vocabulary Use of CLIL and non-CLIL EFL Learners». In: R. M. Jiménez Catalán, & Y. Ruiz de Zarobe (Eds.), Content and language integrated learning: Evidence from research in Europe U.K.: Multilingual Matters. DM:112-130.

Anderson, C., & Freebody, P. (1983). «Reading comprehension and the assessment and acquisition of word knowledge». In: Hutson B. (Ed.), *Advances in reading/language research: A research annual Greenwich*, CT: JAI Press. DM: 231–256.

Canga (2013). «Receptive vocabulary size of secondary Spanish EFL learners». *Revista de Lingüísticas y Lenguas Aplicadas*, vol. VIII: 66-75.

Canga, A. (2015). «Receptive Vocabulary of CLIL and Non-CLIL Primary and Secondary School Learners». *Complutense Journal of English Studies*, vol. XXIII: 59-77.

Coady, J. (1993). «Research on ESL/EFL vocabulary acquisition: Putting it in context». In: T. Huckin, M. Haynes, & J. Coady (Eds.), *Second language reading and vocabulary learning* Norwood, N.J.: Ablex. DM: 3–23.

Jiménez, R. M., & Ruiz de Zarobe, Y. (2009). «The receptive vocabulary of EFL learners in two instructional contexts: CLIL versus non-CLIL learners». In: Jiménez Catalán, R. M. & Ruiz de Zarobe, Y. (Eds.), *Content and language integrated learning: Evidence from research in Europe*. Bristol: Multilingual Matters. DM: 81-93.

Laufer, B. (1986). «Possible changes in attitudes towards vocabulary acquisition research». *IRAL*, vol. XXIV: 69–75.

Laufer, B. (1990). «Why are some words more difficult than others? Some intralexical factors that affect the learning of words». *International Review of Applied Linguistics in Language Teaching*, vol. XXVIII: 293–307.

Laufer, B. (1991). «Knowing a word: What is so difficult about it? ». *English Teachers' Journal*, vol. XLII: 82-86.

Lightbown, P. M., & Spada, N. (1999). *How languages are learned*. Oxford: Oxford University Press.

López, 1995. «Teaching and learning vocabulary: An introduction for English students». *Revista de la Facultad de Educación de Albacete: Ensayos*, vol. X: 35-49.

Meara, P., & Buxton, B. (1987). «An alternative to multiple choice vocabulary tests». *Language Testing*, vol. IV: 142–154.

Meara, P. (1980). «Vocabulary acquisition: a neglected aspect of language learning». *Language Teaching and Linguistics: Abstracts*, vol. XIII (4): 221–246.

Meara, P. (1996a). « The dimensions of lexical competence». In: G. Brown, K. Malmkjaer, & J. Williams (eds.). *Performance and Competence in Second Language Acquisition*. Cambridge: Cambridge University Press. DM: 35-53.

Meara, P. (1996b). *The vocabulary knowledge framework*. Wales University: Vocabulary Acquisition Research Group Virtual Library.

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

Meara, P. (2010). *EFL vocabulary tests* (2nd ed.). Wales University: Swansea Centre for Applied Language Studies.

Milton, J. (2009). *Measuring Second Language Vocabulary Acquisition*. UK: Multilingual Matters.

Milton, J. & Fitzpatrick, T. (2014). *Dimensions of Vocabulary knowledge*. UK: Palgrave Macmillan.

Nation, I. S. P. (1974). «Techniques for teaching vocabulary». *English Teaching Forum*, vol. XII (3): 18–21.

Nation, I. S. P. (1975). «Teaching vocabulary in difficult circumstances». *English Language Teaching Journal*, vol. XXIX: 115–120.

Nation, I. S. P. (1983). «Testing and teaching vocabulary». Guidelines, vol. V (1): 12–25.

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

Nation, I.S.P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nation, I.S.P. (2005). «Teaching Vocabulary. *Asian EFL Journal*, vol. VII (3): 47-54. Retrieved from: http://www.asian-efl-journal.com/sept_05_pn.pdf

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

O'Dell, F. (1997). «Incorporating vocabulary into the syllabus». In: Schmitt, N., and M. J. McCarthy (Eds.) (1997). *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press. DM: 258–278.

Pellicer, A. & Schmitt, N. (2012). «Scoring Yes-No vocabulary tests: Reaction time vs. nonwords approaches». *Language Testing*, vol. XL: 1-21.

Richards, J. C. (1976). «The role of vocabulary teaching». *TESOL Quarterly*, vol. X: 77–89.

Ruiz de Zarobe, Y. (2008). «CLIL and foreign language learning: A longitudinal study in the Basque Country». *International CLIL Research Journal*, vol. I: 60-73.

Schmitt, N. (2008). «Instructed second language vocabulary learning». *Language Teaching Research*, v. XII: 329-363.

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

Schmitt, N., Schmitt, D., & Clapham, C. (2001). «Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test». *Language Testing* vol. XVIII (1): 55-88.

Ur, P. (1996). *A course in language teaching: Practice and theory*. Cambridge: Cambridge University Press.

Xue, G., & Nation, N. (1984). «A university word list». *Language Learning and Communication*, vol. III (2): 215–229.