Creation and validation of a bilingual test to estimate aural and written vocabulary size

MARTÍN AOIZ PINILLOS
Universidad de Navarra

Received: 31 December 2021/ Accepted: 18 June 2022
DOI: 10.30827/portalin.vi38.23606
ISSN paper edition: 1697-7467, ISSN digital edition: 2695-8244

ABSTRACT: Language learners' vocabulary size is a reliable predictor of their success in a second language as it clearly correlates with better performances in the target language (Nation, 2001). Precise estimations of language learners' actual knowledge are paramount to plan language teaching. However, the instruments employed by previous studies for those estimations might present validity and reliability issues that affect their research sensitivity and accuracy. This paper presents a step-by-step account of the creation of an aural and a written version of a bilingual vocabulary test. Answers from 73 adult L1-Spanish students attending English classes were analysed with the Rasch model to determine the best performing items in the test so that the overall reliability of the instrument was enhanced. The final version presents high levels of reliability: .89 for the listening vocabulary test and .82 for the written vocabulary test. Furthermore, descriptive statistics confirm that recognizing the words in their aural form is more challenging than in their written form: participants obtained 10.80% fewer correct answers in the listening vocabulary test. This finding confirms the claim that aural and written vocabulary are two separate dimensions, and impacts on how vocabulary should be taught in L2 classrooms.

Key words: L2 vocabulary, aural vocabulary size, written vocabulary size, vocabulary test, vocabulary teaching.

Creación y Validación de un Test Bilingüe para Calcular el Tamaño del Vocabulario Oral y Escrito

RESUMEN: El tamaño del vocabulario de quienes aprenden una lengua es un predictor fiable de su éxito en la lengua meta porque se correlaciona claramente con mejores rendimientos (Nation, 2001). Las estimaciones precisas del conocimiento real de quienes aprenden idiomas son esenciales para planificar la enseñanza de lenguas. Sin embargo, los instrumentos empleados pueden presentar problemas de validez y fiabilidad que afecten a su sensibilidad y precisión investigadoras. Este artículo describe la creación de una versión oral y escrita de un test bilingüe de vocabulario. Las respuestas de 73 estudiantes adultos de inglés cuya lengua materna era el español fueron analizadas con el modelo Rasch para determinar los elementos del test que mejor se comportaban y mejorar la fiabilidad general del instrumento. La versión final presenta altos niveles de fiabilidad: .89 para la prueba oral de vocabulario y .82 para la prueba escrita. Además, las estadísticas descriptivas confirman que reconocer las palabras en su forma oral supone un reto mayor que hacerlo en su forma escrita porque hubo un 10,80% menos de respuestas correctas en el test oral. Este hallazgo confirma la afirmación de que el vocabulario oral y escrito son dos dimensiones distintas e influye en cómo se debería enseñar el vocabulario en las aulas.

Palabras clave: vocabulario de segunda lengua, tamaño de vocabulario oral, tamaño de vocabulario escrito, test de vocabulario, enseñanza de vocabulario.

1. Introduction

Previous research studies in second language (L2) acquisition have shown the importance of assessing the learner's vocabulary size, as it clearly correlates to better performances, and serves as a predictor of learning success. This positive correlation has been observed in several studies across all language skills. For example, Schmitt et al., (2011) studied the impact of language learners' vocabulary size on their reading comprehension. Similar examples can be found in Andringa et al. (2012) with listening, Crossley et al. (2013) with writing, or Ovtcharov et al. (2006) with speaking. Furthermore, results from these vocabulary assessments might be useful guides for teachers when designing their language courses and sequencing their students' learning (Nation, 2016).

Methodological limitations have been observed in previous research studying L2 vocabulary knowledge, such as the imperfect definition of the counting units (Schmitt et al., 2017), the inadequate use of research instruments to estimate language learners' vocabulary size (van Zeeland, 2014), or the insufficient specificity in both the sample of study participants, and in the selection of the items to be included in the tests, leading to a possible source of measurement error (Ockey & Green, 2020). The more relevant the research instruments and the sample are for the population under study, the more sensitive they are to apprehend the studied phenomena (Mathias, 2010). In this respect, the possibility of creating vocabulary tests based on a limited and more specific corpus might have a positive impact on the overall validity and reliability of such research (Nation, 2016). Moreover, the present study proposes the use of bilingual tests to estimate the vocabulary size of participants sharing a common language, as they might be more sensitive than previous vocabulary tests designed in the target language, as well as facilitate their delivery and replication in different settings (see, for example, Karami, 2012; Nguyen & Nation, 2011; Zhao & Ji, 2018).

The aim of this study is to provide a detailed account of the process followed to design, create, implement, evaluate, and refine a research instrument to estimate the vocabulary size of learners of English as a foreign language (EFL) attending classes in Spain. The specificity of the items to be included in the final version of the vocabulary test was enhanced because they were selected from a corpus which was closely related to the target population. In this thorough account of the steps taken with respect to this corpus-based vocabulary test, particular emphasis will be placed on the importance of using valid and reliable instruments to investigate how a language learner's vocabulary size might predict part of their linguistic performance in a second language.

2. Literature Review

2.1. Unit of measurement in L2 receptive vocabulary knowledge

One of the first issues to address when estimating the vocabulary size of a language learner is what counting unit to employ, e.g., types, lemmas, word families (Nation, 2001). Although defining that measure is essential in any study that implies the use of corpora and their derived vocabulary lists, previous vocabulary studies have imperfectly defined those counting units with no explanations about how they had dealt with occurrences like multiword expressions, homoforms or polysemy (Schmitt et al., 2017).

This study uses word families as the counting unit because is the recommended measure when assessing receptive vocabulary knowledge (Nation, 2016). Researchers assume thus that knowing one or two members of a word family facilitates the receptive knowledge of other members with little effort on the part of the learner (Nation & Webb, 2011). Within this perspective, learners who already know one member of a word family are expected to recognize most of the other members, even if their linguistic proficiency is minimal (Beglar & Nation, 2007).

However, one limitation of studies like the present derives from employing frequency wordlists based on the analysis of corpora, because of the software used in compiling those lists, and in analysing the frequency of the words featured in a given text. The linguistic knowledge of computers in this respect is limited: they can only judge if a given string of characters in the text is different from the next one, and if it has an exact match in any of the entries stored in their databases. Therefore, the use of computers in the analyses of texts might have reduced the concept of 'word' to a match on a list stored in a computer, ignoring the cases of homoforms, polysemy or proper nouns, and neglecting the inclusion of multiword units in the analyses (Cobb, 2013). When dealing with those particular cases, researchers need to assume that some single words might actually be two words, and some phrases could be considered single words. Although there have been attempts to overcome this problem by providing computer systems with new wordlists of proper and compound nouns (Nation, 2012, 2019), and multiword expressions (Martinez & Schmitt, 2012), research about such occurrences is still in its "infancy" (Schmitt et al. 2017, p. 2).

2.2. Research instruments to estimate vocabulary knowledge: validity and reliability

A second set of issues with respect to the assessment of L2-learners' vocabulary size derives from the instruments employed in the estimations. Research has claimed that word frequency is the best measure available to assess the lexical quality of a text (Crossley et al., 2013), and consequently it should guide the selection of words for learners to study (Hazenberg & Hulstijn, 1996). By quantifying the approximate number of words a learner knows, and checking the frequency of the words featured in texts, L2 researchers have attempted to set the minimum vocabulary size necessary to understand different types of texts.

However, some variation exists in the minimal percentage of words a person has to be familiar with for adequate functioning in a second language. Those lexical coverages vary not only depending on the language skill, but the studies also differ in their recommendations for the same skill. Although it might seem small, an increase in the coverage from 95% to 98% of the words featured would imply passing from a vocabulary size of about 3,000 word families to 6,000-7,000 (van Zeeland & Schmitt, 2013). Therefore, percentages of the minimum knowledge necessary to achieve comprehension, or to function adequately in the target L2 have to be highly accurate. This accuracy relies on two separate measurements: research has to be precise when assessing the amount of vocabulary a language user actually has, and when analysing the lexical density of a text in terms of the frequency of its words. The actual validity and reliability of the instruments employed for the estimations of receptive vocabulary knowledge with respect to both the population under study and the final purpose of the assessment might account for part of those differences.

For example, the use of Yes/No tests to assess the learner's vocabulary size (Meara & Miralpeix, 2006) might present validity issues. Three aspects of this type of tests might be criticized. First, the test-takers themselves decide if they know the target words, with some individuals being more lenient than others in their judgements. A second limitation of this type of test refers to the absence of clear criteria about what knowing the target words implies (Nation, 2001); it could be simply knowing that the word exists in the target language, or it could be that they can recall their meaning, or maybe it could mean being able to use it correctly in a sentence. Since the inclusion of nonwords in the test is the only way to control whether the test-taker is accurate in their judgements, an overestimation in the results might occur (van Zeeland, 2014). A final criticism refers specifically to the aural version of the test because it is usually completed on a computer, and the test-takers can play the target word as many times as they wish, and take as long as they want to answer each question (McLean et al., 2015). All these possible flaws in the design of this type of vocabulary tests might have contributed to overestimations of learners' vocabulary size as high as 34.6% (van Zeeland, 2014).

Another strand of criticism with respect to previous research studies refers to employing inadequate instruments to investigate a phenomenon like written vocabulary tests for correlations to L2 listening and speaking (for example, Andringa et al., 2012). In those cases, studies have ignored the claims that learners' ability to recognize words in their written and spoken forms might be different and should be assessed separately (Zhao & Ji, 2018), so that the aural vocabulary knowledge is emphasized as the "primary construct of relevance" (Matthews, 2018, p. 24). However, very few studies have employed listening vocabulary tests to estimate their participants' vocabulary size, despite the higher predictive power of aural vocabulary tests when it comes to L2 listening comprehension (Masrai, 2020).

Moreover, the use of dictation exercises to estimate the aural vocabulary knowledge (for example, Cheng & Matthews, 2018) might raise validity issues as it identifies knowing a word with just being able to recognize its aural form and produce its written form, without having to provide evidence of any link to its meaning. L2 learners with some proficiency in the target language phonology might be able to recognize and transcribe L2 words they have just encountered for the first time, particularly those words that are similar in form to their L1: however, they might fail to make any further sense of them within a broader discourse. Furthermore, in those vocabulary tests, the test-takers have to write the target word within one blank, with other words before and after. Those boundaries are really helpful to the listener to anticipate when to focus their attention on the stream of words, and for how long. As it happens with the aural versions of Yes/No vocabulary tests, the ecological validity of the instrument (Schmuckler, 2001) is negatively affected, as it differs from what a real-life listening situation demands from the listener.

3. Research study

3.1. Research Instruments

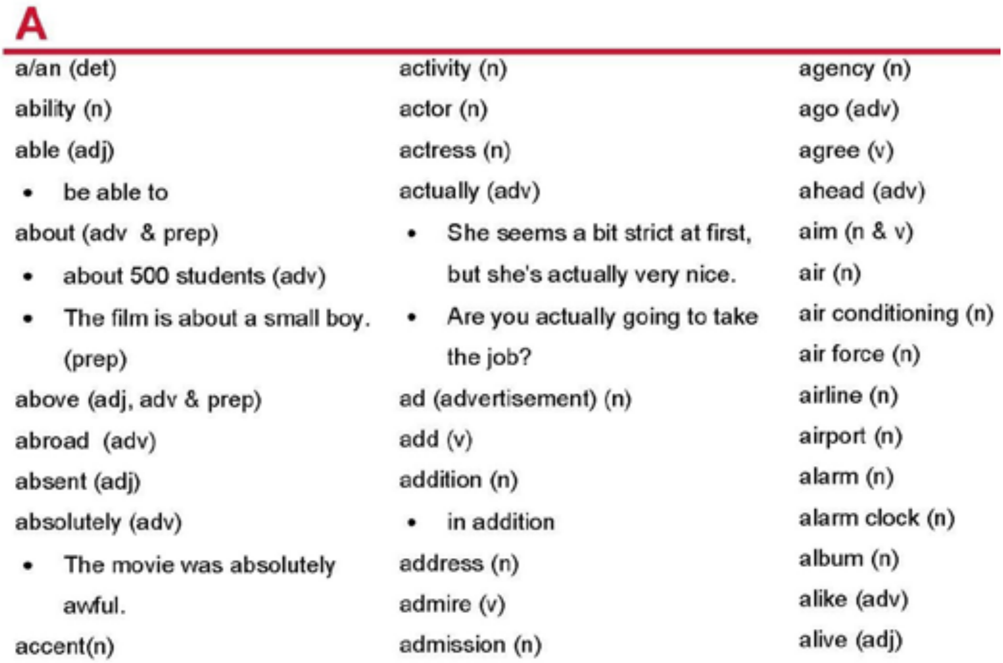
3.1.1. Vocabulary Test – Preliminary Issues

The items for the vocabulary test were selected from the official vocabulary list accompanying the B1-level examination Cambridge English: Preliminary (PET). The counting unit was the types that appear in the PET wordlist (UCLES, 2012) since the test does not specify what level of grouping is intended for the words in that list in terms of types, lemmas, or word families. Consequently, each entry in the PET vocabulary list, including each of the specified meanings in polysemic words like ‘play’, was considered as an independent item for its inclusion in the vocabulary test. This decision enabled the presence of items like ‘improve’ and ‘improvement’, or ‘colour’ as a noun and then as a verb, in the first and unrefined version of the test.

Secondly, multiword expressions like ‘at least’ or ‘bank account’ were excluded from the sampling process because the software employed for text profiling does not include lists compiled by other authors (Cobb, 2019). Only the 1-25k wordlists compiled by Nation (2012, 2019) from the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA) were used in the text profiling analyses. Nevertheless, this use implies an update with respect to previous research because they are the most recent and complete lists of words, based on the “the best corpus of general English in existence” because of its size, balance, and currency (Schmitt & Schmitt, 2012, p. 494).

3.1.2. Cambridge English: Preliminary (PET) – Vocabulary List

The official PET Vocabulary List (UCLES, 2012) was edited before providing the items for the vocabulary tests in the present study. Figure 1 shows part of the original vocabulary list with a total of 2,978 separate entries (i.e., not lemmatized or grouped into word families according to the affixation or derivation they might show). First, all polysemic instances in the list (e.g., ‘play’) were edited to include as many different entries for a word as parts of speech or word meanings had been compiled by its authors, like ‘play the guitar’ and ‘theatre play’. Additionally, multiword expressions without a match on the BNC-COCA vocabulary lists like ‘central heating’ or ‘by mistake’ were excluded from the edited list.



A

a/an (det)

ability (n)

able (adj)

- be able to

about (adv & prep)

- about 500 students (adv)
- The film is about a small boy. (prep)

above (adj, adv & prep)

abroad (adv)

absent (adj)

absolutely (adv)

- The movie was absolutely awful.

accent(n)

activity (n)

actor (n)

actress (n)

actually (adv)

- She seems a bit strict at first, but she's actually very nice.
- Are you actually going to take the job?

ad (advertisement) (n)

add (v)

addition (n)

- in addition

address (n)

admire (v)

admission (n)

agency (n)

ago (adv)

agree (v)

ahead (adv)

aim (n & v)

air (n)

air conditioning (n)

air force (n)

airline (n)

airport (n)

alarm (n)

alarm clock (n)

album (n)

alike (adv)

alive (adj)

Figure 1. PET Vocabulary List (UCLES, 2012, p. 5)

The resulting compilation was analysed with the software Compleat Web VP v.2 (Cobb, 2019), and 3,089 entries found a match among the BNC-COCA 1-25k wordlists. The 124 tokens from the PET list considered ‘off-list’ words were all compounds like ‘birthday’ or ‘bedroom’, except for the word ‘turkey’ and the interjections ‘oh’ and ‘wow’. Table 1 shows a summary of the correspondences of the PET vocabulary list in the 1-25k bands. The last column in the table features the cumulative percentage of tokens in the PET vocabulary test. This figure shows the lexical coverage of a given text that knowing the words up to that band might provide. For example, if a person knows the 3,000 most frequent words in English – according to the frequency lists based on the BNC-COCA corpora – they might be able to understand 84.6% of all words from the PET Vocabulary List.

Table 1. Items in PET Vocabulary List According to Frequency Bands in 1-25k BNC-COCA

FREQ. LEVEL	FAMILIES (%)	TYPES (%)	TOKENS (%)	CUMULATIVE TOKEN %
K-1 Words	961 (44.30)	1,278 (46.99)	1,660 (51.68)	51.68
K-2 Words	634 (29.20)	726 (26.69)	797 (24.80)	76.48
K-3 Words	226 (10.40)	244 (8.97)	261 (8.12)	84.60
K-4 Words	157 (7.20)	163 (5.99)	171 (5.32)	89.92
K-5 Words	89 (4.10)	92 (3.38)	97 (3.02)	92.94
K-6 Words	48 (2.20)	48 (1.76)	49 (1.52)	94.46
K-7 Words	23 (1.10)	23 (0.85)	23 (0.72)	95.18
K-8 Words	15 (0.70)	15 (0.55)	15 (0.47)	95.64
K-9 Words	6 (0.30)	6 (0.22)	6 (0.19)	95.83
K-10 Words	2 (0.07)	2 (0.07)	2 (0.07)	95.89
K-11 Words	3 (0.10)	3 (0.10)	3 (0.10)	95.99
K-12 Words	1 (0.03)	1 (0.03)	1 (0.03)	96.02
K-13 Words				96.02
K-14 Words	3 (0.10)	4 (0.15)	4 (0.15)	96.14
OFF-LIST		114 (4.19)	124 (3.86)	100.00
Total Words	2,168	2,719 (100)	3,213 (100)	100.00

Creation of a Vocabulary Test based on the PET Vocabulary List

Firstly, each of the 3,089 items in the PET list with a match in the 1-25k had its frequency band recorded. Then, the main researcher drew on his experience teaching English to L1-Spanish learners to write down the most suitable and frequent translation into Spanish for each of those terms. Finally, 150 items were randomly selected for their inclusion in a multiple-choice vocabulary test, where they were presented along with four possible translations into Spanish. Figure 2 shows the first items in the Written Vocabulary Test (WVT).

1. TICKET: This **ticket** is perfect

a) cubo
b) entrada
c) factura
d) rama

2. OPERATION: This **operation** is perfect.

a) cuerpo
b) estantería
c) operación
d) pizarra

3. SKIN: This **skin** is perfect.

a) garganta
b) piel
c) pierna
d) uña

4. BORED: They are really **bored**.

a) aburrido
b) ansioso
c) avergonzado
d) decepcionado

5. ASSISTANT: This **assistant** is new.

a) ayudante
b) comerciante
c) representante
d) suplente

6. WEDDING: This **wedding** is perfect

a) boda
b) nacimiento
c) negocio
d) reunión

7. AGAIN: They need it **again.**

a) al revés
b) al mismo tiempo
c) de nuevo
d) depués

8. CLOWN: Thisskin **clown** is perfect.

a) pañuelo
b) payaso
c) peine
d) ternero

9. ICE: This **ice** is perfect.

a) cerradura
b) hielo
c) isla
d) tecla

10. REFUSE: They **refuse** it very often

a) reconocer
b) recuperar
c) reducir
d) rehusar

Figure 2. Written Vocabulary Test

Huang (2010) considered multiple-choice vocabulary tests “simplistic” (p. 4) because they identify vocabulary knowledge with being able to recognize its form and match it to a meaning. However, they are valid research instruments because “the form-meaning link is the first and most essential aspect which must be acquired” when studying L2 vocabulary (Schmitt, 2008, p. 333) because it is “a fundamental first step in gaining control over a particular word” (Cheng & Matthews, 2018, p. 4).

Furthermore, matching the target item to its most suitable translation into Spanish among four options might also contribute to enhance the overall quality of the test. As bilingual vocabulary tests present the words in the target language and the possible answers in the test-takers' L1, they might provide "feasible alternatives to more challenging and time-consuming monolingual tests" (Nguyen & Nation, 2011, p. 86). In monolingual versions of vocabulary size tests, the options are written in the target language in the form of a broad definition, a paraphrase, or a description. Test designers have to be extremely careful in those sentences, and use words that are actually more frequent than the target item. This precaution might be impossible to maintain when testing the knowledge of very frequent words (Beglar & Nation, 2007). Additionally, test-takers with lower levels of proficiency in the target language, might be unfamiliar with some syntactic structures used in those definitions, which might result in testing additional aspects of that language, apart from just their vocabulary size (Nguyen & Nation, 2011). Consequently, when beginners or low-level learners are included among the target population for a study, the use of bilingual tests might be preferable (Levitzky-Aviad & Laufer, 2013), because the respondent's ability to recognize the target items (which should be the focus of vocabulary tests) is not confounded with their ability to read answer options in the L2. A final argument in favour of using translations is that they facilitate test replication with different target languages in other parts of the world, which is crucial in the promotion of transparency and collaboration in research (Abbuhl & Mackey, 2017). In this respect, several research studies into vocabulary testing have already implemented different bilingual versions of vocabulary tests like the present study (Karami, 2012; Nguyen & Nation, 2011; Zhao & Ji, 2018).

The terms for the multiple-choice answers in the test were taken from the same frequency band as the target item, as well as from the same part of speech. Unlike the target items in the test, the actual selection of the four options for each of those vocabulary items was not random but based on the main researcher's intuitions and experience in language teaching and vocabulary testing. The 600 options featured in this version of the vocabulary test (150 target items * 4 options) were used only once.

Each test item was presented both in an isolated manner and within a minimal sentence that only helped determine which part of speech was tested in each case (Figure 2). This manner of presenting the items to the participants intended to show them where to focus their attention, in an attempt to minimise the problems derived from a possible lack of noticing (van Zeeland, 2014). Special care was taken in the selection of the three incorrect options to avoid ambiguity and confusion, and in the writing of contextualising sentences for the target word, so that no additional information about its meaning was revealed. In order to ensure its validity and reliability, a version of the vocabulary test was distributed to five native speakers of English with extensive experience in teaching the language to L1-Spanish learners. They received a version of the vocabulary test with all the target items replaced by a string of characters (XXXXXX), and they were asked both to write down the part of speech they thought each of the items presented in the test, and to supply the correct answer. There were only 10 discrepancies when selecting the part of speech tested in each of the 150 items, which means that in 98.66% of the cases, the teachers coincided in their judgements (5 raters * 150 items = 750 cases). The percentage of coincidence among raters on the part of speech for each vocabulary item was in the range 80-100% (i.e., at least four teachers selected the same part of speech to be assigned to each item). The high level of agreement among those raters showed that the context sentences and the options had been designed correctly, and that the discrepancies in some of their judgements were likely due to the carelessness caused by such a repetitive and taxing task. Furthermore, the contextualising sentences revealed nothing about what answer to choose as none of the five teachers was able to select one option over the rest in any of the questions.

The test was also distributed to six teachers whose L1 was Spanish, and who had years of experience in teaching English to Spanish speakers. They were asked to provide feedback on the clarity of the test with respect to selecting one answer over the other distracters. A total of 99.44% of their answers were correct, and the five incorrect answers referred to different target words, so their mistakes might be attributed to carelessness. Furthermore, none of the teachers raised concerns about the possible ambiguity of any of the options, or the incorrectness of any of the translations. These results clearly confirmed both the validity of the Spanish translations for each item in the test and the overall unambiguity of the options to select the correct answer for each question.

When all 150 items in the test, their contextualising sentences and the four options for each question were revised by those 11 experienced teachers, the listening vocabulary test was created. A recording studio was booked, and a native English speaker was asked to read out each of the 150 words and their context sentences in the test, as well as the introductory instructions and examples (Figure 2). The only indication he received was to read the text as clearly and naturally as possible, without attempting to conceal his idiolinguistic features (i.e., accent, prosody, intonation, etc.). Once the recording session finished, the raw audio file was edited with the software Audacity®, and the questions in the test were separated 5 seconds from each other, sufficient for the test-takers to read the four options and select the correct one in each case (van Zeeland, 2014).

3.2. Data collection

Adult students attending English B1-level classes at the Official Language School and the Institute of Modern Languages of the University of Navarra in Pamplona (Spain) were invited to participate in the study. An online version of the test was created on Google Forms® and sent to all the students who had accepted to take part in the study. Participants were told to answer all the questions, including those that were totally unknown to them, to enhance the ecological validity of the test (Schmuckler, 2001). In real-life situations language users sometimes might have to make tentative guesses at the meaning of the unknown words they encounter. Furthermore, participants were unaware of the fact that both the listening vocabulary test (LVT) and the written vocabulary test (WVT) presented the same items first orally and then in writing, but once in the WVT, they were asked not to change any of their answers in the previous section, so that they were entirely based on the aural input from the recording.

Although the design of the present study made it impossible to set controls for practice-of-order effects, a subsequent analysis suggested that neither the first items in the test were more difficult to answer correctly because the participants had no experience with the test format; nor were the last items in the test more difficult because the test-takers were tired. Table 2 shows the scores in the LVT and the WVT, divided in thirds with 50 items each, and in the order they were delivered, first the aural and then the written version of the test. The mean scores for the three thirds in the WVT are higher than their counterparts in the

LVT, although test-takers answered those questions later. Similarly, the second third in both the LVT and the WVT presents lower mean scores than the last third in each of the tests.

Table 2. Results in Consecutive Thirds of Items in Vocabulary Tests

	LISTENING VOCABULARY TEST				WRITTEN VOCABULARY TEST			
	Total	*Third 1*	*Third 2*	*Third 3*	*Total*	*Third 1*	*Third 2*	*Third 3*
	150 items	*50 items*	*50 items*	*50 items*	*150 items*	*50 items*	*50 items*	*50 items*
Mean score	131.08	44.08	43.40	43.60	138.81	47.41	45.33	46.07
SD	9.10	4.07	2.84	3.76	7.22	2.40	2.46	3.13
% Correct	87.39%	88.16%	86.79%	87.21%	92.54%	94.82%	90.66%	92.14%

The data for mean scores and percentage of correct answers featured in Table 2 show that the lack of experience to answer the first items in a test, or the fatigue and boredom caused by such a demanding and repetitive task had no impact on the participants' success to select the right answer. The mean scores for the three thirds in each test show that they depend both on the difficulty of the items and on the test modality, not on the order or sequence of the items: the second third is more difficult than the other two thirds, and the LVT is more difficult than the WVT.

3.3. Data analysis

3.3.1. Descriptive Statistics: Reliability, Separation

The data from the participants (n = 73) were imported onto the program Winsteps® (Linacre, 2012, 2019) to be analysed using the Rasch model, which implies accepting explicitly the interval nature of data, because counts "cannot replace measurement as it is known in the physical sciences" (Bond & Fox, 2015, p. 6). This model converts raw scores, which are equivalent to counting, into linear and reproducible measurement. By means of a probabilistic match, it conjointly analyses two factors that affect the performance in a test, the person's ability, and the item difficulty. In practical terms, the Rasch model offers the researcher a single unit of measurement called 'logit', which enables the comparison of items and persons on the same scale, as well as the comparison of different samples of people, or different items related to the same observed trait.

The first steps in the data analysis focused on the reliability of the test, given its particular importance in applied linguistics (Hatch & Lazaraton, 1991). Overall, both participants and items showed higher separation and reliability indices in the listening than in the written vocabulary test. The reliability index reported by the Rasch model is similar to more traditional ones in test theory like KR-20 or Cronbach's alpha (Linacre, 2012). The closer the values are to 1, the more internally consistent is the measure. However, those traditional reliability indices are considered to overstate "the reliability of the test-independent, generalizable measures the test is intended to imply. For inference beyond the test, Rasch reliability is more conservative and less misleading" (Linacre, 1997, p. 581), because it avoids misinterpreting raw scores as linear measures.

The main reason for the differences in the reliability and separation indices between the LVT and the WVT might lie in the lower number of persons or items with extreme scores. In the LVT all participants chose the correct answer for 33 items, but no test-taker got a perfect score. Consequently, the standard error of measurement (SEM) was higher in the WVT because two people answered all 150 questions in the WVT correctly, and for 60 items (40%) all test-takers selected the right option. A subsequent data quality analysis was undertaken to increase reliability and separation in the tests by eliminating items with perfect or nearly perfect scores because they conveyed too little information about the participants' performance. At the same time, in order to allow comparisons between the participants' aural and written vocabulary size, items were excluded only when there were minimal differences in scores from one test modality to the other. Table 3 shows the number of items with perfect scores and their corresponding separation and reliability indices, depending on how many items are analysed. Among the 150 items from the original dataset, 29 presented perfect scores in the LVT, and 43 in the WVT. When items with either perfect scores in both tests or 72/73 correct answers in one test and perfect scores in the other were excluded from the analysis, a total of 121 items remained in both the LVT and the WVT. The item reliability and separation indices increased for both tests because they had fewer items with perfect scores, whereas the person reliability and separation varied minimally. The selection of the best performing items continued until items had perfect scores in one test and a minimum of 68/73 correct answers in the other, so all differences between the scores for the excluded items in the two versions of the test were smaller than 7%.

Table 3. Perfect Scores Depending on Number of Items Considered and Corresponding Values for Separation and Reliability (Expressed in Logits)

	COUNT PERFECT SCORES		PERSON SEPARATION		PERSON RELIABILITY		ITEM SEPARATION		ITEM RELIABILITY	
	LVT	*WVT*	*LVT*	*WVT*	*LVT*	*WVT*	*LVT*	*WVT*	*LVT*	*WVT*
150 Items	33	60	2.46	1.93	.86	.79	1.64	1.09	.73	.54
121 Items	4	31	2.46	1.93	.86	.79	2.21	1.36	.83	.65
105 Items	0	19	2.44	1.94	.86	.79	2.59	1.55	.87	.71
97 Items	0	11	2.41	1.94	.85	.79	2.66	1.70	.88	.74
91 Items	0	11	2.40	1.94	.85	.79	2.77	1.75	.88	.75
81 Items	0	1	2.34	1.94	.85	.79	2.87	2.12	.89	.82

Table 3 clearly shows how excluding the items where all or nearly all participants found the correct answer enhances the separation and reliability indices. As items with perfect scores are dismissed, the separation between participants slightly decreases or stays the same. On the other hand, when fewer items are considered in the analysis, the separation among them logically increases because the range of difficulty – from the easiest to the most difficult item – might be slightly shorter, but the number of individuals covering that distance is certainly smaller. The person reliability index is barely affected by the number of items in the analysis because the number of participants remains the same (*N* = 73); whereas the item reliability dramatically increases, in particular in the written version of the test.

3.3.2. Descriptive Statistics: Item Difficulty

As the percentage of correct responses in the listening and written vocabulary test were 77.93% and 86.35% respectively, we can conclude that the aural version of the test was more difficult. Furthermore, the test format (aural or written) might be responsible for the higher difficulty of the LVT because the target words in both vocabulary tests were the same. Table 4 presents the main descriptive statistics: for the LVT, the mean person ability was 1.79 logits (*SD* =.82), whereas participants showed a mean ability of 2.77 logits (*SD* = 1.03) in the WVT.

Table 4. Descriptive Statistics for the Listening and Written Vocabulary Test

	PERSON					ITEM				
	RAW SCORES			LOGITS		RAW SCORES			LOGITS	
	Count	*Mean*	*SD*	*Mean*	*SD*	*Count*	*Mean*	*SD*	*Mean*	*SD*
LVT	81	63.10	8.30	1.79	.82	73	56.9	12.90	.00	1.20
WVT	81	69.90	7.10	2.77	1.03	73	63.0	10.80	-.04	1.32

In the LVT, one item (L51, 'shut', 3.26 logits) was clearly the most difficult one, followed by items L23 ('wide', 2.46 logits) and item L90 ('handle', 2.39 logits). The easiest items in the LVT were L35 ('switch', -2.81 logits) and L134 ('glove', -2.89 logits). For the WVT, the most difficult item was W51 ('shut', 3.18 logits), followed by W62 ('have', 3.11 logits). The easiest word in that test was W88 ('item', -3.27 logits), followed by a group of 12 items (e.g., 'pig' or 'creature') with a difficulty of -2.04 logits each. It is worth mentioning that although all participants in the written test chose the correct translation for item W88 ('item'), it was not removed from the analysis because only 64 participants chose the correct option for its counterpart in the LVT (L88, -0.43 logits), which might imply an important difference across the two versions of the vocabulary test.

With respect to the test-takers' abilities, in the WVT only one participant showed an ability slightly inferior to the average difficulty of its items. In other words, that person had an overall chance of getting a correct answer for any item in the test a bit lower than 50%. On the other hand, for the LVT, although no participant showed abilities lower than 0 logits, their mean ability with respect to that test was inferior to the ones shown by the same participants in the WVT. In both tests performance values were distributed along a continuum, so different levels of ability (persons) and difficulty (items) could be observed.

4. Discussion

This investigation aimed at creating a valid and reliable test to estimate both the aural and the written receptive vocabulary size of English language learners whose first language was Spanish. The official vocabulary list accompanying a standardized language test (*Cambridge English: Preliminary*) was used as the corpus from which to select the items in the tests. The answers from 73 study participants were analysed and a refinement process was undertaken to determine the best performing items in the tests. Items conveying too little information about the participants' performance were excluded, which led to an increase in both reliability and separation indices.

The final version of the written and listening vocabulary test showed high levels of reliability (.82 and .89, respectively), in line with what previous studies had presented. Bilingual versions of written vocabulary size tests had showed reliabilities ranging from .78 (Cheng & Matthews, 2018) to .96 (Nguyen & Nation, 2011; Karami, 2012). In aural vocabulary tests, the Listening Vocabulary Size Test (McLean et al., 2015) showed a reliability of .98, probably because it was tested on almost three times more participants than the present study (*N* = 214 vs *N* = 73). In the case of the Aural Vocabulary Test employed by Matthews (2018), reliability indices ranged from .78 to .81, depending on the band of frequency.

The present study was a partial replication of the investigation carried out by McLean et al. (2015), as it created a listening vocabulary test from the items in a written vocabulary test. Furthermore, both tests presented a multiple-choice bilingual format with the four options for each of the target words translated into the participants L1, so that their validity and reliability indices could increase (McLean et al., 2015). However, the present study delivered both modalities of the vocabulary tests to the same participants to enable comparisons with respect to their ability to recognise words either in their aural or written form. Although employing a multiple-choice format and translating the options into the test-takers' L1 might also raise concern about its validity and reliability (Huang, 2010), the alleged cognitive load that might derive from the use of translations in this kind of vocabulary tests has not been detected in qualitative analyses (McLean et al., 2015). Furthermore, this format has proved to be a valid and highly reliable method for vocabulary size estimations (Silva & Otwinowska, 2019).

Although the main objective of the present study was to create a reliable instrument to estimate the participants' aural and written vocabulary sizes, it is worth mentioning the differences in the results depending on the type of test, i.e., aural or written. The percentage of correct answers increased by 10.80% from the LVT to the WVT (Table 4). This result might be in line with what Masrai (2020) found out when he compared the aural and written vocabulary size of L2-English learners. However, the present study has been the first one to estimate both dimensions of vocabulary on the same subjects and at the same moment in time, which increases its validity and reliability. This disparity in the performance shown by the same learners with respect to the same items, which depended mainly on the test modality, is a solid argument in favour of considering learners' aural and written vocabulary size as two separate dimensions that need to be assessed separately (Zhao & Ji, 2018). Furthermore, it implies challenging the approach used in previous studies with respect to L2 vocabulary estimations because they might have employed inadequate research instruments.

5. Conclusion

This paper has presented the first aural and written vocabulary test specifically designed for Spanish learners of English. This article is a step-by-step guide to the creation of bilingual vocabulary tests and it shows that creating valid, reliable, and sensitive instruments to estimate the vocabulary size of language learners is a feasible process that language teachers can undertake. Being aware of the ease and feasibility of that task and familiarise oneself with the steps in the process of designing a vocabulary test might be of particular interest in highly specific fields of language teaching where such instruments might not be available.

Additionally, this study has produced evidence to support the claims that the language learner’s aural and written vocabulary size are two different dimensions that need to be assessed separately, and that language learners know more words in their written than in their aural form. Based on both findings, L2 teachers might need to challenge some of the assumptions made by previous research into L2 vocabulary, particularly the estimations of the minimal vocabulary necessary to understand aural texts in a second language. A possible consequence is the inclusion in L2 methodologies of new approaches to teaching vocabulary as language learners’ aural vocabulary size seems to be relatively smaller. Nevertheless, further research is necessary to confirm the existence of these two vocabulary dimensions, aural and written, and to estimate the possible differences.

6. References

Abbuhl, R., & Mackey, A. (2017). Second language acquisition research methods. In King, K. A., Lai, Y. J., & May, S. (Eds.). *Research methods in language and education* (3rd edition). (pp. 183-193). https://doi.org/10.1007/978-3-319-02249-9

Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of success in native and non□native listening comprehension: An individual differences approach. *Language Learning, 62*(Suppl. 2), 49–78. https://doi.org/10.1111/j.1467-9922.2012.00706.x

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

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: fundamental measurement in the human sciences* (3rd ed.). Routledge.

Cheng, J., & Matthews, J. (2018). The relationship between three measures of L2 vocabulary knowledge and L2 listening and reading. *Language Testing, 35*(1), 3-25. https://doi.org/10.1177/0265532216676851

Cobb, T. (2013). Frequency 2.0: Incorporating homoforms and multiword units in pedagogical frequency lists. In Bardel, C., Lindqvist, C., & Laufer, B. (Eds.) *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis*, (pp. 79-108). EUROSLA-the European Second Language Association. Retrieved from https://www.eurosla.org/

Cobb, T. (2019). *Compleat Web VP v.2* [computer program]. Accessed on 16 Jan 2019 at https://www.lextutor.ca/cgi-bin/range/texts/index.pl

Crossley, S. A., Cobb, T., & McNamara, D. S. (2013). Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. *System, 41*(4), 965-981. https://doi.org/10.1016/j.system.2013.08.002

Hatch, E. M., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Heinle & Heinle Publishers.

Hazenberg, S., & Hulstijn, J. H. (1996). Defining a minimal receptive second language vocabulary for non-native university students: An empirical investigation. *Applied Linguistics, 17*(2), 145-163. https://doi.org/10.1093/applin/17.2.145

Huang, H. T. (2010). *How does second language vocabulary grow over time? A multi-methodological study of incremental vocabulary knowledge development* (Doctoral dissertation, University of Hawai'i).

Karami, H. (2012). The development and validation of a bilingual version of the Vocabulary Size Test. *RELC Journal, 43*(1), 53-67. https://doi.org/10.1177/0033688212439359

Levitzky-Aviad, T., & Laufer, B. (2013). Lexical properties in the writing of foreign language learners over eight years of study: Single words and collocations. In Bardel, C., Lindqvist, C., & Laufer, B. (Eds.) *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis*, (pp. 127-148). EUROSLA-the European Second Language Association. Retrieved from https://www.eurosla.org/

Linacre, J. M. (1997). KR-20 / Cronbach alpha or Rasch person reliability: Which tells the "truth"? *Rasch Measurement Transactions, 11*(3), 580-581

Linacre, J. M. (2012). *A user's guide to Winsteps Ministeps Rasch-model computer programs* [version 3.74.0]. Retrieved from http://www.winsteps.com/index.htm

Linacre, J. M. (2012, 2019). *Winsteps® Rasch Measurement, version 4.4.3*. [Computer software] Downloaded from http://www.winsteps.com

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics, 33*(3), 299-320. https://doi.org/10.1093/applin/ams010

Masrai, A. (2020). Exploring the impact of individual differences in aural vocabulary knowledge, written vocabulary knowledge and working memory capacity on explaining L2 learners' listening comprehension. *Applied Linguistics Review, 11*(3), 423-447. https://doi.org/10.1515/applirev-2018-0106

Mathias, C.W. (2010). Sensitivity. In Salkind, N. J. (Ed.). (2010). *Encyclopedia of research design (Vol. 3),* (pp. 1337-1338). Sage.

Matthews, J. (2018). Vocabulary for listening: Emerging evidence for high and mid-frequency vocabulary knowledge. *System, 72*, 23-36. https://doi.org/10.1016/j.system.2017.10.005

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

Meara, P. M., & Miralpeix, I. (2006). *Y_Lex: The Swansea advanced vocabulary levels test. v2. 05.* Lognostics.

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

Nation, I. S. P. (2012, 2019). *The BNC/COCA word family lists (17 September 2012)*. Unpublished paper. [online] Retrieved from http://www.victoria.ac.nz/lals/about/staff/paul-nation

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. John Benjamins. https://doi.org/10.1093/applin/amx052

Nation, I. S. P., & Webb, S. A. (2011). *Researching and analyzing vocabulary*. Heinle, Cengage Learning.

Nguyen, L.T.C., & Nation, P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. *RELC journal, 42*(1), 86-99. https://doi.org/10.1177/0033688210390264

Ockey, G. J., & Green, B. A. (Eds.). (2020). *Another generation of fundamental considerations* *in language assessment: A festschrift in honor of Lyle F. Bachman*. Springer Nature.

Ovtcharov, V., Cobb, T., & Halter, R. (2006). La richesse lexicale des productions orales: mesure fiable du niveau de compétence langagière. *Canadian Modern Language Review*, *63*(1), 107-125.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research, 12*(3), 329-363. https://doi.org/10.1177/1362168808089921

Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2017). How much vocabulary is needed to use English? Replication of van Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007). *Language Teaching, 50*(2), 212-226. https://doi.org/10.1017/S0261444815000075

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*(1), 26-43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Schmitt, N., Schmitt, D. (2012). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, 47*(4), 484-503. https://doi.org/10.1017/S0261444812000018

Schmuckler, M. A. (2001). What is ecological validity? A dimensional analysis. *Infancy, 2*(4), 419-436. https://doi.org/10.1207/s15327078in0204_02

Silva, B. B., & Otwinowska, A. (2019). VST as a reliable academic placement tool despite cognate inflation effects. *English for Specific Purposes, 54*, 35-49. https://doi.org/10.1016/j.esp.2018.12.001

UCLES (2012). *Cambridge English: Preliminary Wordlist*. Retrieved from https://www.cambridgeenglish.org/Images/84669-pet-vocabulary-list.pdf

UCLES (2021). *B1 Preliminary and B1 Preliminary for Schools Vocabulary List.* Retrieved from https://www.cambridgeenglish.org/images/506887-b1-Preliminary-2020-vocabulary-list.pdf

van Zeeland, H. (2014). *Second language vocabulary knowledge in and from listening* (Doctoral dissertation, University of Nottingham).

van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics, 34*(4), 457-479. https://doi.org/10.1093/applin/ams074

Zhao, P., & Ji, X. (2018). Validation of the Mandarin version of the Vocabulary Size Test. *RELC Journal, 49*(3), 308-321. https://doi.org/10.1177/0033688216639761