Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 12, No. 2, October 2022
ISSN 2476-5880

A Bilingual Version of the Vocabulary Size Test for Spanish Speakers

Benjamin Carcamo[1]*

Received: November 2021 Accepted: April 2022

Abstract
The objective of this study was to validate a bilingual Spanish-English version of the Vocabulary Size Test (VST) considering its potential use as a discriminator between learners in terms of language competence. This version was designed based on the two forms available on one of the creators' websites as well as considering practices recommended regarding the elimination of cognates and loans. A one-way ANOVA test was used to confirm the test's capacity to discriminate among learners of different linguistic competence. Additionally, Principal Axis Factoring (PAF) was conducted to revise the existence of only one underlying variable. As a result of this study, a VST version for Spanish speakers consisting of 9 vocabulary frequency levels is shared. This version is in line with validation standards put forward in previous research. It is expected that this instrument will help future studies that seek to measure Spanish speakers' competence in English as a foreign or second language without having to deal with the interference of other intervening factors.

Keywords: learning; testing; validity; vocabulary

1. Introduction

Assessing the level of vocabulary of an ESL student is a challenging task for which varied instruments have been designed. One of the most widely used tests for vocabulary measurement is the Vocabulary Size Test (VST). This multiple-choice instrument is capable of measuring the receptive knowledge of 14,000-word families frequently used in the English language. In terms of format, the instrument offers the examinee a word without any context and four alternatives from which the examinee must recognize the one with the closest meaning (Nation & Beglar, 2007).

Even though there are several studies that have made use of this instrument (Masrai, 2019; Şen & Kuleli, 2015), there are still questions to be addressed regarding its implementation. One of these questions is whether or not the monolingual version of the test should be used with learners of English as a second or foreign language. The assumption underlying this possible problem is that if students sit for the VST in English, vocabulary knowledge is already a variable that could affect the understanding of the instrument. For example, an examinee could know the meaning of a word, but still be unable to identify the alternative that resembles it the most because of a lack of vocabulary. In this respect, Elgort

[1] Facultad de Comunicaciones, Universidad de Las Américas, Sede Providencia, Manuel Montt 948, Santiago, Chile. Email: benjamin.carcamo@pucv.cl

(2013) has asserted that, instead, bilingual versions of the VST are ideal for intermediate and beginner ESL/EFL learners.

Regarding the translation of this document, several versions have been created. Among these languages, one can find bilingual versions for Mandarin (Zhao & Ji, 2016), Russian (Elgort, 2013), Vietnamese (Nguyen & Nation, 2011), Japanese (Derrah & Rowe, 2015), and Persian (Karami, 2012; Karami et al., 2017). Nonetheless, the literature review shows that there appears to be no bilingual version for Spanish speakers even though Spanish is one of the most spoken languages around the world and the importance of vocabulary research in ESL (Reynolds & Teng, 2021; Zabolotna, et al., 2021).

The objective of this study is to propose a bilingual version of the VST for Spanish speakers that is able to determine their linguistic competence in English as a Foreign Language. To do this, the article is divided as follows. First, it offers information about how vocabulary is usually assessed and how vocabulary tests are validated. Then, the article shares the method followed to validate the VST version for Spanish speakers. After this, the results are presented as well as the discussion of these findings. Finally, conclusions are drawn from the results.

2. Literature Review

2.1 How to measure vocabulary knowledge

Vocabulary knowledge is a multidimensional construct. In fact, knowing a word may entail knowing multiple dimensions such as pronunciation, collocations, and synonyms, among others (Schmitt, 2000; Schmitt, Schmitt & Clapham, 2001). Nevertheless, vocabulary tests tend to focus exclusively on the measurement of the ability to recognize the prototypical meaning of words because of the predictive power this dimension has been proved to have over variables such as reading comprehension (Laufer & Aviad-Levitsky, 2017).

Two instruments stand out for their popularity and efficiency when measuring the extent of vocabulary knowledge: The Levels Test (Nation, 1983; Schmitt, Schmitt & Clapham, 2001) and the Vocabulary Size Test (VST; Nation & Beglar, 2007). The Levels Test is a diagnostic evaluation whose main purpose is to help the learner detect areas of weakness. In order to do this, the Levels test includes slices of words from different sets, such as the Academic Word List (AWL), the most frequent words, and the least frequent words. Therefore, if the student gets low scores in a particular kind of word, for example, in some of the items related to the AWL, then he or she knows what to review. Regarding the format, in the Levels test the student receives in each item a set of words which they have to match with their corresponding definition.

On the other hand, the VST has as its main purpose to determine the proficiency students have by providing a reliable measure of the vocabulary size the learner possesses from the 1st 1000 to the 14th 1000-word families of English (Nation & Beglar, 2007). Consequently, the uses of the VST results are very different from those of the Levels Test. For example, the VST can help the teacher or researcher to map out the growth of learners' vocabularies by using the test with the students at different points in time. In addition, the VST results are useful to estimate whether or not the learner has the vocabulary required to understand certain texts based on the complexity of the text and the word families that the learner has mastered. Regarding the format, the VST is a multiple-choice test. The words included in the test are

taken from 14 of the 10000 wordlists of the British National Corpus. BNC wordlists are based on word families, which at the time the authors considered to be one of the best indicators of vocabulary knowledge. The main belief was that passing a minimum level of competence, second language learners should be able to handle word formation skills through the manipulation of affixes (Nation & Beglar, 2007).

From this perspective, word families are understood as the root plus inflectional and derivational forms. Thus, a word such as *use* would be from the same family as the words *useless* or *reuse*. That is, while the root is an independent form (*work*), it can be considered that somebody that knows the meaning of the root should understand the meaning of its associated family of words. Even though this belief has lost validity in recent years, the test itself is still widely used in many interventions (Stoeckel, McLean, & Nation, 2020).

The use of the classic monolingual version of this test has been validated in different instances. Based on these successful applications, the creators have highlighted the following positive aspects of the instrument (Beglar, 2010):

- Can be used with learners of different levels of competence provided minor adjustments are made, such as the length of the test.
- Measures what it should measure, which in this case is receptive vocabulary knowledge.
- Behaves as expected, distinguishing different levels of competence.
- Performs consistently and reliably even when circumstances change.
- Is easy to correct and interpret.
- Has clear and unique items.
- Provides relevant results to determine the different levels of L2 English competence even if taking only a 70-item version of the test.

Suffice to say, even though the frequency of occurrence is most of the time deemed as the strongest factor affecting vocabulary knowledge, there are others affecting the likelihood of a word being known, such as whether the word is a cognate in the learners' L1, learners' opportunities to have encountered the word before, the morphological transparency of the word, among others (Hashimoto & Egbert, 2019; Laufer, 1997). For this reason, recently, researchers have been exploring the accuracy with which the VST measures vocabulary knowledge for some of the purposes it has been used. Stoeckel, McLean, and Nation (2020) investigated some weaknesses of the Levels as well as the Size tests. These researchers have identified potential issues with the use of meaning-recognition formats, the assumption that knowing a word implied the knowledge of a family word, and the use of random samples in the construction of vocabulary tests. Although these kinds of arguments seem to put into question the effectiveness of the tests, the authors indicate they still can be used to rank and group learners based on the results they offer.

2.2 Validation process for the bilingual versions of the VST

The VST test has originated bilingual versions in different languages, such as Russian, Korean, Mandarin, Japanese and Vietnamese versions (Karami, 2012). The bilingual versions are characterized by having the alternatives to the questions in the test-takers' language. The following example shows how an item in Spanish would look.

3. shoe: Where is your <shoe>?
a. la persona que cuida a alguien b. el objeto donde se guarda el dinero
c. el objeto que usas para escribir d. el objeto que usas en tu pie

As for the validation of the bilingual versions, there are some differences in the processes the researchers went through. An example of this is the research conducted by Nguyen and Nation (2011) who created a bilingual version of the VST for Vietnamese speakers. For the validation process, the authors started by translating the four possible alternatives of each item into the L1 of the target examinees. It was important for the authors to avoid word-by-word translation and, instead, look for the closest equivalents of the English vocabulary in the Vietnamese language. Specifically, the translation procedure was performed by two English teachers who were native speakers of Vietnamese. First, one teacher translated all the alternatives into Vietnamese, and then the other teacher read and corrected the translation. The translations were then reviewed by both to ensure the accuracy and intelligibility of all the items. Likewise, translations of the interpretation of the scores were made, which helped explain the instrument to the subjects that participated in the process.

To validate the bilingual version, the test was administered to 62 Vietnamese students who were in their third year of study at a university in Vietnam. From this sample, three groups of a similar number with different linguistic competence were generated: A beginner group (N = 20), an intermediate level group (N = 21), and an advanced group (N = 21). As input to perform this classification, the averages of the four courses taken in the previous semester were used: Reading, writing, audio translation, and translation theory. For the validation of the VST, the test results were expected to distinguish between three levels of competence: beginner, intermediate and advanced. Before answering the test, students were allowed to ask questions. As for the application process, they could take all the time they needed to answer the questions in the instrument. The main statistical procedure to validate the instrument was the use of a one-way ANOVA that distinguished the levels of English of the students considering the complete test as well as the first seven levels. This was a successful validation and one of its main strengths was the way in which the translation of the items was conducted.

Karami (2012) designed and validated the bilingual version of the VST in Persian. This test was calibrated with 190 Iranian English learners (Beginner Group = 91, Intermediate Group = 77, Advanced Group = 22). The test design process followed the procedure established by Nguyen and Nation (2011). The distractors were translated from English to Persian using Persian equivalents whenever possible. Even though loanwords were identified in the VST as potentially problematic for the validity of the instrument, it was decided to keep them in the bilingual version. The generated bilingual version of the VST was sent to five native Persian speakers who had graduated from English teaching programs. After editing the test based on

their comments, the instrument was sent back to the panel of experts in order to amend the last details. As an extra step, the designed instrument was piloted with a group of 10 students, who were later interviewed informally about their experience with the test. The aim of this application was to find ambiguities in the translations and confirm if these students could answer less frequent word items.

The test was validated with two statistical procedures. First, a factor analysis of the 14 frequency levels was done by means of the Principal Axis Factoring (PAF) extraction method. This analysis showed that only one factor was identified with an Eigenvalue over 1, which explained 55% of the variance. This confirms that the test measures only one construct. Secondly, the researcher used a one-way ANOVA to show that the differences were significant between the three means. That is, it is confirmed that the Persian bilingual VST is able to differentiate between different levels of competence of the subjects.

More recently, a bilingual version of the VST for Japanese test-takers has been developed (Derrah & Rowe, 2015). The validation of this version was done with 43 students who studied in two different contexts. A first group (N = 27) attended a university for women and were in the third year of a communication program with English courses, TOEIC/TOEFL or both. The other group (N = 16) attended a private high school and had several hours of weekly English in Japanese and one hour of communication in English with a native speaker.

The procedure for the design of the bilingual test was not specified in the article. The authors limited themselves to establishing that the options were translated into Japanese and that the application was made through a website to which the participants had access by entering a password. A Rasch model which calibrated scales of the ability of those who took the test and the difficulty of the items was implemented by means of the Winsteps software.

For the Chinese bilingual version, Zhao and Ji (2016) designed a shortened version of the VST in Mandarin based on the work of Wang and Fan (2011). This version was piloted with 177 Chinese students in 3 different stages of linguistic development. The first group consisted of 35 students majoring in English. The second group was second-year students in an electronic program. The third group included 50 students who were in the first year of an international business program. In addition, this study made use of the construct validation framework proposed by Messick (1995).

The one-way ANOVA test was statistically significant. In fact, all differences were significant among groups. The reliability of the items was .97, which is similar to the one obtained in previous validations (Japanese: a=92, Persian: a=.97, Vietnamese: a = .96 (Derrah & Rowe, 2015; Karami, 2012; Nguyen and Nation, 2011). It should be noted that Zhao and Ji (2016) also distinguished loanwords as problematic in the correct evaluation of the less frequent levels in their bilingual version. In light of this, it can be asserted that the inclusion of these kinds of words seems to be a problematic factor that has been distinguished by the majority of the authors in the reviewed studies, but rarely has it been addressed.

After examining the validations of bilingual versions, certain strengths and weaknesses can be identified. The main commonality in terms of strengths has to do with the process of translation of the items. The use of a committee of experts (Karami, 2012) and the consideration of more than one translator (Nguyen & Nation, 2011) have been proved to be effective. A second commonality is related to the statistical procedures used to assert that the instrument is

valid. One-way ANOVA and factor analysis appear to be the most common techniques employed. The one-way ANOVA test is used to confirm that the bilingual version is able to discriminate among different levels of English competence as the monolingual version does. That is, if a learner scores lowly in the VST, it means that they have a low level of English and if they score highly, they have a high level of English. Additionally, a few of the validations make use of factor analysis to confirm that, after the translation takes place, there is still just one underlying variable explaining most of the variance.

Regarding the weaknesses, there is a clear inconsistency between the discussion of limitations of the studies and the practices incorporated regarding cognates and loanwords. Researchers who have designed bilingual versions of the VST have agreed on the fact that the presence of cognates may affect the validity of the test, especially when they are in the last few levels of frequency (Karami, 2012; Nguyen & Nation, 2011; Zhao & Ji, 2016). Likewise, researchers have detected loanwords as potentially negative to the effectiveness of the instrument. This fact is supported by psycholinguistic empirical studies which have shown, for example, that cognates are acquired more easily and memorized better independently of their frequency (De Groot & Keijzer, 2000; Liu, 2008). Therefore, one can conclude that the elimination of cognates, as well as loanwords, should increase the validity of bilingual versions of the VST.

3. Method

3.1 Research Questions

Considering that the main purpose of this research was to design an English-Spanish bilingual version of the Vocabulary Size Test (VST for Spanish speakers) which served as one of its main purposes to group learners according to L2 linguistic competence based on their results. The research questions put forward were the following:

RQ1: Does the bilingual version of the test distinguish learners from different proficiency levels?

RQ2: Does a single construct underlie the bilingual version of the VST for Spanish speakers?

RQ3: Does the level of difficulty of the test increase as the level of frequency of the words decreases?

3.2 Participants

The sample consisted of 100 undergraduates studying English enrolled in the TEFL, Translation, or Interpretation programs at a university located in Chile. Participants received no credits or any type of incentive for their participation in the study. These participants were contacted by one of the researchers who visited the compulsory EFL courses in which they were enrolled (Beginner, Intermediate, or Advanced) to extend an invitation to participate in the study. The syllabus of each of the three courses the researcher visited declared it focused on different levels of English according to the Common European Framework of Reference for Languages (CEFR). The Beginner course developed English language skills at an A2 level, the Intermediate at a B2 level, and the Advanced course at a C1 level. Table 1 summarizes the distribution of the total sample.

ISSN 2476-5880

Table 1.
Sample of the study

Level of English	N
Beginner	31
Intermediate	30
Advanced	39
Total	100

3.3 Validation

Two main factors have been found important in the validation process of bilingual versions of the VST. First, many previous investigations have mentioned a lack of control over cognates and loans, which has been put forward as potentially harmful to the results and interpretation of the scores obtained in the test. Secondly, the use of ANOVA and factor analysis to confirm the predictive validity of the test as well as its construct validity.

As for the first issue, the present study ensured the elimination of cognates and loanwords during the translation stage. For the identification of cognates, the proposal of Beinborn, Zesch, and Gurevych (2014) was taken. These researchers consider as cognates all the pairs of words that are sufficiently similar so that the student can link them, regardless of their root. Loanwords were understood as those lexical pieces that enter one language from another, resulting in the host language accommodating the lexical item grammatically and phonologically according to its own sets of rules (Hoffer, 2005). Two EFL teachers who are native speakers of Spanish went through the item list identifying cognates and loanwords separately. Later they compared the lexical items they found and agreed on the main list of cognates and loanwords which had to be replaced.

Regarding the second highlighted aspect, the use of an ANOVA test was selected as the best statistical procedure to determine the capacity the VST had to discriminate among different competence levels. A Principal Axis Factoring (PAF) with the Direct Oblimin rotation was used to ensure the designed instrument complied with construct validity by making sure one underlying variable, equivalent to vocabulary knowledge, was still present after the translation of the instrument.

3.4 Procedure

The first step to design the instrument was to download the versions A and B of the VST from the official website (https://www.victoria.ac.nz/lals/about/staff/paul-nation). After doing so, the A version was used as the basis for the test. Two Spanish native speakers who are certified teachers of English as a Foreign Language revised the instrument in order to find cognates and loans. Disagreements were solved through a meeting that took place after both teachers checked the instrument independently. A total of 29 lexical items were detected.

The next step was replacing these lexical items with others from Version B of the instrument. In order to do so, it was made sure that the items were of the same level of frequency. For instance, the loan *cyborg* (n° 93 in Version A) was replaced by *bylaw* (n° 97 in Version B). Likewise, the cognate *monologue* (n° 42 in Version A) was replaced by *sizzle* (n° 43 in Version B). Executing this procedure helps making sure that the knowledge of the L2

vocabulary is the only one being measured since a student with little vocabulary knowledge would still know the meaning of a cognate which appears as a very low frequency word, such as *cyborg*.

Taking into account previous studies in which using less than the total 14 levels of frequency have shown satisfactory results (Nguyen & Nation, 2011 [7 levels]; Zhao y Ji, 2016 [8 levels]), this study has taken only the first 10 levels of the test for the design of the bilingual version. The 100 items selected were translated by two EFL teachers separately. After they both translated the instrument, they met to discuss and agree on the items that were translated differently. In the cases in which there was no agreement, a third EFL teacher was consulted in order to settle the disagreement. Once the final version of the test was ready, the 100 subjects sat for the instrument in their respective English courses. The reliability of the test was estimated at .881 with the 100 items. To determine the capacity of the instrument to discriminate among the 3 English levels, an ANOVA test was used.

Once the instrument and purpose of the study was explained to them, the students took the test in their regular classrooms. As suggested by the designers of the VST, students were prompted to make informed guesses and avoid rushing since they had plenty of time to answer the questions.

4. Results and Discussion

After confirming the assumptions of normality and homogeneity of variance were met, descriptive statistics were used to analyze the results of the sample.

Table 2.
Descriptive statistics

Level of English	N	Mean (score)	St. Deviation
Beginner	31	43.387	11.723
Intermediate	30	50.833	10.706
Advanced	39	56.341	7.973

The results of the one-way ANOVA test were statistically significant with a large effect size ($F(2, 97) = 16.166$, $p = .000$; $\eta^2 = .249$). Bonferroni was used as post hoc test. This post hoc test revealed the significant differences, which occurred among different groups: Between beginner and intermediate (-7.466, $p = .013$), beginner and advanced (-13.613, $p = .000$), and intermediate and advanced (-6.1667, $p = .037$). All of the differences were statistically significant.

Then, ten independent one-way ANOVA tests were run (one with each of the frequency levels), the results of the first 9 levels showed the capacity to discriminate among the three levels of the group. The following table shows the results per level.

ISSN 2476-5880

Table 3.
Analyses for scores on each of the levels

Level	df	F	p	η^2
1	(2, 97)	7.347	.001	0.132
2	(2, 97)	7.904	.001	0.141
3	(2, 97)	4.01	.021	0.076
4	(2, 97)	15.953	.000	0.2475
5	(2, 97)	5.094	.008	0.095
6	(2, 97)	5.893	.004	0.108
7	(2, 97)	6.348	.003	0.116
8	(2, 97)	10.792	.000	0.182
9	(2, 97)	8.793	.000	0.153
10	(2, 97)	2.267	.109	0.045

These results suggest that it would be more effective and time-efficient to use only the first 9 levels of the Vocabulary Size Test for Spanish Speakers since level 10 is not able to discriminate among different levels of English. The reliability of the test with 90 questions is still high since it reaches .878. A possible interpretation of the poor results in the low-frequency bands is the possibility that students guess the meanings which weakens the interpretations of the scores in those bands (Gyllstad, McLean, & Stewart, 2020; Stoeckel, McNeal, & Nation, 2020).

To answer the second research question, Principal Axis Factoring (PAF) was used after inspecting the value of the Kaiser-Meyer-Oklin (KMO) and the statistical significance of Bartlett's Test of Sphericity. Direct Oblimin rotation was used with the data. First, Exploratory Factor Analysis (EFA) was conducted with the aim of extracting Eigenvalues over 1. The results of the analysis revealed 2 factors as appropriate for the model. Nonetheless, considering the ANOVA results, as well as the literature and previous research the Confirmatory Factor Analysis (CFA), was run indicating one factor to be extracted, but now considering only the nine levels that were proved to be statistically significant in the previous one-way ANOVA.

The KMO was .872, which is well over the recommended .6 (Netemeyer, Bearden, & Sharma, 2003). Bartlett's Test of Sphericity was statistically significant ($X^2(36) = 357.963$, p=.000). The result of this analysis showed only one factor underlying the test in the examination of the scree plot. This factor explained over 49.62% of the variance. Taking this as well as the previous results, the final VST for Speakers of Spanish has only 9 levels. Figure 1 shows the scree plot.

ISSN 2476-5880

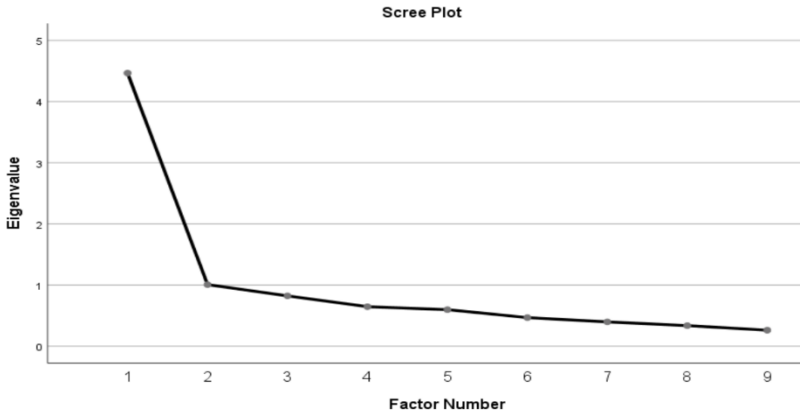

Figure 1. Scree plot

Finally, in order to tackle the final research question, the scores of the subjects were examined. This task was undertaken to determine if there was an increasing level of difficulty in the bilingual version of the test as it occurs with previously validated ones. Table 4 displays the means for each level of frequency in the validation process of the Vocabulary Size Test for Spanish Speakers.

Table 4.
Score means based on frequency level

Level	*M*	SD	Minimum	Maximum
First	8.5	.937	6	10
Second	7.6	1.75	2	10
Third	6.5	1.56	3	10
Fourth	5.1	2.13	0	9
Fifth	5.5	1.86	1	10
Sixth	4.9	1.69	1	9
Seventh	3.6	1.73	0	8
Eighth	3.4	1.69	0	7
Ninth	3.1	1.58	0	6
Tenth	2.7	1.53	0	7

It is evident that there is a clear progression from high scores to lowest ones although this does not occur in strict order across the frequency levels. As it can be noticed level four shows a mean lower than the one of the fifth level. Nevertheless, it is clear that participants perform better on the frequency levels of easier difficulty and worse in the harder ones. This as well as the minimum and maximum scores are in line with the expected results based on other studies (Karami, 2012; Karami et al., 2017). As for the total score, neither ceiling nor floor effects can be reported considering the lowest score was of 19 pts. and the highest was of 78 pts. Taking the results of the test as a whole the mean for the examinees was 50.93.

5. Discussion

Validating tests is of paramount importance in the fields of second language acquisition and TESOL. One of the main reasons why this procedure is important is because of the inferences that are related to the results of a test (Brown & Abeywickrama, 2010). For example, the low scores of a vocabulary size test might lead educators to implement actions to help students with their vocabulary growth. However, were the test to be invalid, there would be no way to know with certainty whether or not these actions are indeed helping the students or even if these low scores represented the vocabulary size of the students in the first place.

In the case of bilingual tests, there is a clear responsibility with both original designers of the test and the students who will sit for them. When an already valid test is translated, there is a responsibility of the original designers to preserve the validity of the test, which implies that this new version keeps the underlying factor or factors (Karami et al., 2017) and that its inferences can still be made (Nguyen & Nation, 2011). The latter, in particular, pertains to the responsibility of the students. If the translated version does not allow educators to make the same inferences made with the original test, then this version is not useful for either the academia or the classrooms.

The present study shares with the community a bilingual version of the VST for Spanish speakers. In order to generate this version, three techniques were used to ensure the validity of the test. First of all, cognates and loanwords were replaced considering previous experiences in which these lexical items were deemed to be problematic in preserving the validity of bilingual versions of the VST (Karami, 2012; Nguyen & Nation, 2011; Zhao y Ji, 2016). Secondly, the use of a one-way ANOVA helped determine that the inferences that can be made by the bilingual version generated in this study are similar to the ones made by the original. Specifically, it can be stated that the VST discriminates among Spanish-speaking learners of different proficiency in English. Finally, the use of factor analysis has helped us determine that the translated version has preserved the underlying factor. The only weakness that can be mentioned is that compared with other bilingual versions the variance explained by this factor is lower. Whereas the present bilingual version explains 49.62% of the variance, the Persian explains up to 53% (Karami, et al. 2017).

Overall, the data analysis of this study advocates for the use of the resulting instrument consisting of the first 9 levels of the VST. This was confirmed by the one-way ANOVA which corroborated that these 9 levels were enough to discriminate among the three proficiency groups. The use of less than the total levels available for the test is in accordance with the belief that low and intermediate proficiency learners do not necessarily have to sit all levels of the test. Examples of this are in Beglar (2010) who used the first four and the first eight levels in different cases, Elgort (2013) who used seven levels, and Zhao and Ji (2016) who used the first eight levels. In addition, in the projections of other validation studies, researchers have prompted the need to validate shorter versions of the VST for Spanish speakers (Nguyen and Nation, 2011).

When examining the results in light of the six suggested questions proposed by Karami (2012), the validity of the bilingual version generated in this research can be asserted with certainty. First, there are no ceiling or floor effects in the test. Second, the test does have a high level of reliability (.878). Third, as it has been shown through factor analysis, there is only one

construct underlying the test. Fourth, the ANOVA test has revealed that the test is in fact able to discriminate among different proficiency levels. Fifth, there is a clear order in terms of difficulty among frequency levels. Finally, based on the results of the test it was determined that the application of the ninety-questions test was the best way to assess the vocabulary knowledge of English of Spanish speakers.

6. Conclusion

Assessment has far-reaching consequences for the people involved in it, such as students, teachers, and stakeholders (Mohammadkhah, et al., 2022). Therefore, it is of paramount importance to design valid tests that help teachers obtain results that truly represent students' abilities. The present article has shown the design and validation procedures for a version of the Vocabulary Size Test made for Spanish speakers. This research has considered pointers offered by researchers who have undertaken this task in other languages, such as Russian, Japanese, Persian, Vietnamese, and Chinese. The main outcome of the research is a validated version of the VST for Spanish speakers.

In order to avoid interference from knowledge other than English vocabulary knowledge, cognates and loans were eliminated from the original instrument, as it had been suggested in previous research. The generated instrument showed high reliability as well as the capacity to discriminate among different levels of English. Moreover, factor analysis showed that the VST for Spanish Speakers had only one underlying construct, which is what can be appreciated in other bilingual versions that have been validated as well as in the original test. This factor clearly represents vocabulary knowledge, which is the variable that the instrument attempts to measure.

When reflecting on the limitations of the study, one comes to mind. Although the procedures followed for the validation of the VST were appropriate, one extra measure might have been taken: the piloting of the items with a preliminary group of learners. Before the implementation of the test, piloting with some learners could have shed light on mistakes related to translation. Instead of doing this, in the study, it has opted for a third EFL teacher who helped solve disagreements among the researchers translating the items. Doing both could help have a more holistic perspective on each of the items translated.

Two projections can be made based on the results of this study and what is being done in the field. First of all, future cross-validation of the performance of the bilingual version and the monolingual version. After the validation of other bilingual versions, they have been cross-validated with the monolingual version in order to compare their performance side to side. This can help specify the nuances of what is the optimal context for using one or the other version (Elgort, 2013; Karami, et al., 2017) and it can also open up avenues to help develop teacher's assessment literacy levels which have been shown to be low (Ahmadi, Ghaffary, & Shafaghi, 2022). Secondly, this bilingual version of the VST can be applied in different contexts with Spanish speakers learning English as a Foreign language. Its application and use are of the utmost importance to corroborate the usefulness of this version of the instrument as well as its stability when used with Spanish speakers of other countries. Validation, after all, is a process that must be constant and persist throughout time (Messick, 1989) while considering the purpose for which the test is given, how is going to be interpreted, and its construct validity.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Ahmadi, S., Ghaffary, S., & Shafaghi, M. (2022). Examining teacher assessment literacy and instructional improvement of Iranian high school teachers on various fields of study. *International Journal of Language Testing*, *12*(1), 1-25.

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

Beinborn, L., Zesch, T. & Gurevych, I. (2014). Readability for foreign language learning. *International Journal of Applied Linguistics*, *165*(2), 136-162.

Brown, H. & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices*. New York: Pearson Education.

Cohen, J. (1969). *Statistical power analysis for the behavioural sciences*. New York: Academic Press.

De Groot, A. M. B. (2011). *Language and cognition in bilinguals and multilinguals: An introduction.* New York, NY: Psychology Press.

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

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

Gyllstad, H., McLean, S., & Stewart, J. (2020). Using confidence intervals to determine adequate item sample sizes for vocabulary tests: An essential but overlooked practice. *Language Testing*, *38*(4), 558-579.

Hashimoto, B. & Egbert, J. (2019). More than frequency? Exploring predictors of word difficulty for second language learners. *Language Learning*, *69*(4), 839-872.

Hoffer, B. (2005). Language borrowing and the indices of adaptability and receptivity. *Intercultural Communication Studies, XIV*(2), 53-72.

Karami, H. (2012). The development and validation of a bilingual version of the vocabulary size test. *RELC Journal, 43*(1), 53-67. doi: 10.1177/0033688212439359

Karami, H., Nejad, M., Nourzadeh, S. & Shirazi, M. (2017). Validation of a bilingual version of the vocabulary size test: Comparison with the monolingual version. *International Journal of Bilingual Education and Bilingualism*, 23(4), 368-370. doi: 10.1080/13670050.2017.1391744

Laufer, B. & Aviad-Levitzky, T. (2017). What type of vocabulary knowledge predicts reading comprehension: word meaning recall or word meaning recognition? *The Modern Language Journal, 101*(4), 729-741.

Laufer, B. (1997). What's in a word that makes it hard or easy? Intra-lexical factors affecting the difficulty of vocabulary acquisition. In McCarthy, M. & Schmitt, M (1997), *Vocabulary description, acquisition and pedagogy* (pp. 140-155). Cambridge: Cambridge University Press.

Liu, J. (2008). L1 use in L2 vocabulary learning: Facilitator or barrier. *International Education Studies, 1*(2), 65-69.

Masrai, A. (2019). Vocabulary and reading comprehension revisited: Evidence for High-, Mid-, and Low-frequency vocabulary knowledge. *SAGE Open, 9*(2), 1-13. doi: 10.1177/2158244019845182

Messick, S. (1989). Validity. In R. L. Linn (Ed.) *Educational Measurement* (3rd. ed) (pp. 13-104). New York: American Council on Education and Macmillan.

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

Mohammadkhah, E., Kiany, G. R., Tajeddin, Z., & ShayesteFar, P. (2022). Teachers' conceptions of language assessment: Affective and theoretical knowledge dimensions of language assessment literacy model. *International Journal of Language Testing*, 12(1), 82-102.

Nation, P. & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*, 9-13. doi: 10.1177/0033688210390264

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

Netemeyer, R., Bearden, W., & Sharma, S. (2003). *Scaling procedures: Issues and applications*. London: Sage.

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

Reynolds, B.L., & Teng, M.F. (2021). Incidental and informal vocabulary learning: Introduction to the special issue. *TESOL Journal, 12*(4), 1-7. doi: 10.1002/tesj.642

Schmitt, N., Schmitt, D. & Clapham, C. (2001). Developing and exploring the behavior of two new versions of the vocabulary levels test. *Language Testing, 18*, 55-88. doi: 10.1177/026553220101800103

Şen, Y., & Kuleli, M. (2015). The effect of Vocabulary Size and vocabulary depth on reading in EFL context. *Procedia Social and Behavioral Sciences, 199*, 555-562. doi: 10.1016/j.sbspro.2015.07.546

Stoeckel, T., McLean, S., & Nation, P. (2020). Limitations of size and levels test of written receptive vocabulary knowledge. *Studies in Second Language Acquisition*, *43*(1), 1-23. doi: 10.1017/S027226312000025X

Wang, Y. & Fan, Y. (2011). Validity study on Chinese version of VST. *English Abroad, 11*, 310-311.

Zhao, P. & Ji, X. (2016). Validation of the Mandarin version of the Vocabulary Size Test. *RELC Journal*. doi:10.1177/0033688216639761

Zabolotna, O., Zagoruiko, L., Pachenko, I., & Plotnikov, Y. (2021). Teaching English vocabulary online: Is the screen a barrier? *Advanced Education, 17*, 57-64.