Development of an instrument for measuring Spanish vocabulary knowledge

How Keat Khong
Universiti Kuala Lumpur

Muhammad Kamarul Kabilan
Univiersiti Sains Malaysia & Universitas Negeri Malang

Received: 2023-07-30 / Accepted: 2023-11-07
DOI: *https://doi.org/10.30827/portalin.vi41.25406*
ISSN paper edition: 1697-7467, ISSN digital edition: 2695-8244

ABSTRACT: This study addresses the need for a robust vocabulary assessment instrument to offer reliable estimates of vocabulary knowledge. Current measures often lack validation evidence and mainly serve instructional purposes. Therefore, this study presents the development of a valid, reliable assessment tool for research and evaluation. Adhering to Read and Chapelle's (2001) framework, we carefully selected context-specific resources to enhance the validity and reliability. Our results affirm the instrument's strong psychometric properties, demonstrating good validity (S-CVI/Ave: pre-test = .97; post-test = .98) and reliability (KR-20: pre-test = .883; post-test = .932). Moreover, the instrument displays sensitivity to various vocabulary instructional approaches [Kruskal-Wallis test, H (2) = 11.609, p = .003 from pre-test to post-test; H (2) = 12.434, p = .002 from pre-test to delayed post-test]. This study promotes vocabulary assessment literacy among language educators and researchers, ensuring the integrity of instrument development and enhancing language learning and research. By documenting the creation and validation of this assessment tool, our study contributes a valuable resource that enables educators and researchers to improve language education in accordance with contemporary measurement practices.
Keywords: instrument, vocabulary test, vocabulary knowledge, validity, reliability

Desarrollo de un instrumento para medir el conocimiento del vocabulario del español

RESUMEN: Este trabajo aborda la necesidad de contar con un instrumento sólido de evaluación del vocabulario que ofrezca estimaciones fiables del conocimiento léxico. Las medidas actuales a menudo carecen de evidencia de validación y se utilizan principalmente con fines educativos. Por tanto, este trabajo presenta el desarrollo de un instrumento de evaluación válido y fiable para la investigación y la evaluación. Siguiendo el marco de Read y Chapelle (2001), seleccionamos cuidadosamente recursos específicos del contexto para mejorar la validez y la fiabilidad. Nuestros resultados confirman las sólidas propiedades psicométricas del instrumento, demostrando una buena validez (IVC-escala/promedio: pre-test = 0,97; pos-test = 0,98) y fiabilidad (KR-20: pre-test = 0,883; post-test = 0,932). Además, el instrumento muestra sensibilidad a diversas estrategias instructivas de vocabulario [prueba de Kruskal-Wallis, H (2) = 11,609, p = 0,003 de pre-test a post-test; H (2) = 12,434, p = 0,002 de pre-test a post-test diferido]. Este trabajo promueve la alfabetización en la evaluación del vocabulario entre los profesores y los investigadores de idiomas, garantizando la integridad del desarrollo del instrumento y mejorando el aprendizaje y la investigación lingüística.

Al documentar la creación y validación de este instrumento de evaluación, nuestro estudio contribuye con un recurso valioso que permite a profesores e investigadores mejorar la educación lingüística de acuerdo con las prácticas de medición contemporáneas.
Palabras clave: instrumento, test de vocabulario, conocimiento del vocabulario, validez, fiabilidad

1. Introduction

Reliable estimates of vocabulary knowledge (VK) of any foreign language (FL) learners offer teachers with useful information for improving related learning performance. Nonetheless, existing vocabulary measures (or instruments) are predominantly used for instructional purposes (e.g., placement, achievement, and proficiency of learners). To our knowledge, little attention has been paid to the development of standard measures for research and evaluation uses. Concurring with Read and Chapelle (2001), language teachers avoid vocabulary assessment may be due to "the only measures available seem to be irrelevant to their needs" (p. 22). Moreover, the existing debate in the vocabulary literature reflects a lack of validation evidence. For instance, Aoiz Pinillos (2022) observes various methodological limitations in previous research including poor definition of the units of measurement, inadequate use of research instruments, lack of specificity in research subject and in test item selection.

Therefore, the objective of this study is to address the gap by providing a description of the development of valid and reliable vocabulary measures to evaluate the effects of different vocabulary instructions tailored for beginners. We believe that there are teachers and researchers whose teaching and learning contexts disallow them to simply adopt any available instruments which are usually more suitable for measuring VK of higher proficient learners. In short, this study adhered to the framework for vocabulary assessment (Read & Chapelle, 2001) whereby the instrument was created using resources which were closely related to the research subject and context to enhance the validity and reliability. Results of the field study and preliminary evidence for the validity and reliability of the instrument were included.

2. Literature review

2.1. Conceptualising vocabulary knowledge

Literature reveals that VK, word, and measurement are often inseparable. Initially, terms like 'lexical unit' (Cruse, 1986) and 'base form' (Sinclair & Renouf, 1988) were used to conceptualise the notion of word. While these terms are persuasive, they lack the ability to indicate what exactly constitutes a word, and they do pose potential problems in subsequent vocabulary measurement. For pragmatic reasons, other alternative terms, or units of word measurement, including token, type, lemma, and word family have been foregrounded (Nation & Hunston, 2013; Read, 2000; Schmitt, 2010). All these indicate that in conceptualising the notion of a word, researchers are encouraged to consider the use of multiple elements of word measurement and make explicit which term they adopt in their studies, so that the readers will be adequately informed what word is being studied and thus allowing comparability of vocabulary studies.

In our research, the term 'lemma' was adopted to measure learners' VK. Lemma consists of a headword or a base (or dictionary) form plus its set of inflected forms, all of which are under the same part of speech. This adoption was based on the following criteria. Firstly, lemma reflects the research objectives that seek to investigate the effects of two interventions on different VK aspects. This criterion is consistent with Kremmel (2018), Schmitt (2010) and Sánchez Gutiérrez (2021) that lemma is a more reliable unit of counting words with a transparent definition across the field. Secondly, 'type' was not adopted because numerous homographs were identified during the vocabulary selection process. For example, 'cocina' can mean 'cook' as a verb or 'kitchen' as a noun. In contrast, counting 'token' might render an overestimation of learners' VK. Lastly, owing to the Spanish verbal flexemes and other irregularities, lemmas seem to be a better option than "word family" (see Read, 2000, p. 19–20 for a comprehensive example) to accommodate the proficiency level of the research subject (they were attending CEFR A1.1 Spanish classes). In sum, this study intends to adopt a more transparent conception (lemma) to assimilate the study into the current research and practice in relation to VK among beginner FL learners.

2.2. Operationalising the construct

During the literature review, we notice that there is yet a generally accepted method to operationalise the construct of VK, hence we adopted a components approach comprising Nation's (2001) framework to measure learners' VK because: (a) Nation's framework reflects the multidimensional nature of VK. While it is argued that this framework is vast and may pose some problems in real-world application, Kremmel (2018) advocates that this framework is "the most thorough and useful view of lexical knowledge to date" (p. 17) and Schmitt (2010) supports that it is the most comprehensive and widely referred framework by current vocabulary scholars; (b) by breaking down the VK into separate components, it makes the measurement practically manageable. By doing so, it allows exploring the relationships and order of acquisition of different components including how they interact and affect each other in the development of the four language skills? How these components contribute to the system knowledge and stored wholes? These questions begin to receive attention in the literature.

Nevertheless, it is nearly impossible to measure all nine VK components in one single test due to the complexity of the construct. This explains why many studies only measured a single component. While single component vocabulary studies are informative, Kremmel (2018) rightfully points out that "a balanced measure that accounts for breadth and some depth would thus seem an important contribution to the field" (p. 20). Consequently, based on the level of the research subject, this study responded to the call and measured four aspects: (a) written form; (b) form and meaning, (c) concept and referents, and (d) associations, to reflect their VK after the interventions.

2.2.1. Written form (orthography)

Knowing the written form of a word includes being able to recognize the word during reading and to produce the word while communicating. In Spanish, there are similar forms

which can lead to potential confusions, for example, 'casa-caso-cosa' (house-case-thing) are similar forms with different meanings, while 'esta-este-esto' (this-this-this; these demonstratives have different grammatical functions in Spanish) are similar forms with similar meanings. Schmitt (2010) warns against side-lining the orthographic/phonological mastery of word form. Hence, learners of this study are required to identify and supply the written form of the target language (TL) through listening to reflect their knowledge of this aspect.

2.2.2. Form and meaning

Knowing a word is typically regarded as knowing the word form and its meaning. However, Nation (2020) emphasises the connection between the two because it is possible that learners know only the form of a word but not the actual conceptual meaning, or learners are familiar with the form and have the corresponding concept, but they fail to connect the two. For example, learners might be aware of the form 'libro' (book), but they might confuse it with 'libra' or 'library' of English. Hence, learners of this study are required to supply the meaning in first language (L1, Malay) or second language (L2, English) through the form given in the TL and supply the translation in the TL through the meaning given in L1/L2 to reflect their knowledge of this aspect.

2.2.3. Concept and referents

Some words consist of one conceptual meaning with only one sense, for instance, 'junio' (June), and some carry two or more unrelated meanings (i.e., 'banco' can mean a bench as in a long seat or a bank as in a financial institution). A polysemous word refers to a word that has one core meaning with various peripheral senses (Verspoor & Lowie, 2003), like 'clase' (class). However, this study focuses on the core meaning of a word, though it may not be the most established as literal and central, but it has a clearly concrete referent that alludes to everyday concepts which are consistent with the course syllabus. Hence, learners of this study are required to select the conceptual meaning in L1/L2 through the form given in the TL and select the answer in the TL which matches the conceptual image given to reflect their knowledge of this aspect.

2.2.4. Associations

When learners learn new words and new concepts, they also expand and consolidate their VK simultaneously by establishing associations among known words and concepts. According to Ma and Lee (2019), there are three main groups of word associations in the mental lexicon: (a) paradigmatic associations, (b) syntagmatic associations, and (c) form-based associations. Considering the proficiency level of the learners, this study focuses only on the paradigmatic associations. Hence, learners of this study are required to select the associated theme in L1/L2 through the form given in the TL and select the answer in the TL among the associated cues given to reflect their knowledge of this aspect.

3. Present Study

The present study belongs to a larger mixed-methods research that aims to investigate the influence of two interventions on Spanish language learners' vocabulary acquisition. To reiterate, this study intends to provide a description of the development of valid and reliable vocabulary measures to evaluate the effects of different vocabulary instructions tailored for beginners using a more restricted corpus. We believe that this study will make the development of instruments possible for researchers and practitioners who could not resort to any established measures to estimate learners' VK in their research that involves interventions.

The participants (research subject) were 19 to 25 years old (M = 21.03, SD = 1.69) technical majors enrolled in a Spanish language course at a university in Malaysia. Initially, 76 students participated in the study: experimental group received intervention A (n = 31), comparison group received intervention B (n = 22) and control group without intervention (n = 23). This sample size was in accordance with Onwuegbuzie and Collins (2007) who proposed 21 participants per group for one-tailed hypotheses for moderate effect sizes with .80 statistical power at the 5% level of significance. However, attrition occurred throughout the 13-week research period and the final sample was 37, experimental group (n = 15), comparison group (n = 12) and control group (n = 10). As part of other ethical considerations, all participants were given equal opportunity to access to all learning materials after the research via the university e-learning website.

3.1. Instruments

3.1.1. Scope of content

The scope of content is defined as the intended vocabulary to be learned. As the Spanish manual was the primary reference for the learners, it was used to define the vocabulary content for the study. Standard corpus (e.g., CORPES XXI) was not referred because the objective of the study was to assess the VK of the first-time learners of Spanish after the interventions wherein the content was consistent with the course syllabus. Hence, the manual was analysed using WordSmith Tools 7.0. The analysis indicated that there were 1,234 lemmas and from these lemmas, a total of 30 themes were identified. Six themes were removed due to practical reasons. For instance, it would not be conducive to learn 30 themes in a short period of time (e.g., during the single-session intervention), and it would be difficult to adequately assess all the 30 themes in one vocabulary test. Therefore, themes like 'greeting', 'farewell' and 'basic communication' were not included because they would be frequently practiced in daily class and via social media communications while 'occupation', 'country' and 'language' were also excluded because their constituent elements would be too vast. As a result, 24 themes remained. To complete the mandatory 28 sessions for the other intervention, four themes were duplicated including 'number', 'family', 'colour' and 'verb'.

Subsequently, the word selection for the 24 themes was based on the following two criteria: the learning needs and the word frequency. For students' learning goals and needs, listening test and oral test were referred because they assessed students' overall vocabulary compared to single topic assessments. For word frequency, Davies's (2006) frequency dic-

tionary of Spanish was used as a word rank frequency reference considering its robustness and representativeness. Also, in line with Nation and Hunston (2013) in that high-frequency words should receive most attention in any vocabulary learning program as well as the need to learn a range of themes which involve mid-frequency (e.g., temporal reference like 'lunes', Monday) and low-frequency vocabulary (e.g., cultural reference like 'paella', a Spanish rice dish) from the course content perspective, a ratio of 6:3:1 for high, mid, low-frequency words was adopted. This ratio helped ensure that the vocabulary learning would occur in the most useful sequence for learners to gain the most benefit from the vocabulary selected. Based on these premises, a total of 140 words (28 x 5) were selected from the manual for the design and development of the research interventions and test instrument. The complete target word list can be found at https://bit.ly/3AJTE0N.

3.1.2. Design and development

Three vocabulary tests were created to measure the four VK aspects individually and collectively before (pre-test), a week after (post-test) and a month after (delayed post-test) the interventions. These tests were considered as criterion-referenced tests (CRTs) in comparison with the norm-referenced tests (see Brown, 1996). They were designed and developed following the four development phases proposed by Benson and Clark (1983, cited in Creswell, 2012).

All tests used the same format. They comprised four sections which assessed the four VK aspects. The sequence of the test items was consistent with Laufer et al. (2004) and Webb (2005, 2007) in that all the recall exercises would be completed before the recognition exercises to avoid "learning effect" (Webb, 2005, p. 39). For the recall exercises which involved orthography and form-meaning aspects, a translation format was used. Meanwhile, for the recognition exercises which involved concept and referents, and associations aspects, a multiple-choice format was used. The semantically unrelated distractors were taken from the 140-word list. Hence, if a learner could select the correct answer based on their understanding of the distractors (i.e., eliminating the wrong possibilities) but the target word itself, the test item could still be considered as valid to the extent that the distractors also constituted the target words to be learned.

In terms of the test length, each section comprised ten items and this amounted to 40 items per test, which equal to 40 (or 60 if the distractors were included) target words learned. On the one hand, Nation and Hunston (2013) suggest that a reliable vocabulary test should have at least 30 items, and on the other hand, the more items a test has, the more time will be needed to complete it, therefore this can discourage learners from completing the test correctly. Therefore, 40 items for a test were deemed appropriate. The four sections were further divided into eight parts (Part 1 to Part 8), and each consisted of five items (with maximum mark = 5) where learners were required to respond either receptively or productively. This could ascertain which test score influenced which type of VK to assure that the eventual gain was clear. Considering the level of the learners under study, the duration of each test was set as 30 minutes based on the pretesting results. Examples of validated test item for each part are presented in Figure 1.

Part 1: Productive recall of orthography
The students listen to each target word pronounced twice and write it correctly in the space given.
1. _____

Part 2: Receptive recall of orthography (after second revision)
The students listen to each short sentence pronounced twice and then circle the right option. The distractors were created to resemble the target word both phonetically and orthographically.
6. (a) tocar (b) doga (c) drogar (d) torca

Part 3: Productive recall of form-meaning
The students complete the target word based on the English/Malay meaning given as in a translation test.
*11. man (lelaki) h*_____

Part 4: Receptive recall of form-meaning
The students translate each target word into English/Malay.
16. quién _____

Part 5: Productive recognition of associations
The students select one option which does not fit with the other four associated words.
21. (a) usted (b) ella (c) tú (d) yo (e) en

Part 6: Productive recognition of concept and referents
The students match the photo with the most suitable word from the list.
26. <Photo> □ cama □ cuadro □ ordenador □ salón □ vestido

Part 7: Receptive recognition of associations
The students select the most suitable option that can be associated with the target word.
31. divertido (a) character (b) furniture (c) food (d) vehicle

Part 8: Receptive recognition of concept and referents
The students select the correct option that corresponds to the core meaning of the target word.
36. él (a) he (b) tall (c) when (d) uncle

Figure 1. *Examples of Test Item (Part 1 to Part 8)*

In terms of scoring, two systems with different levels of precision were used. According to Nakata and Webb (2016), scoring responses at two levels of sensitivity are more likely to provide a more accurate assessment of vocabulary learning. Firstly, a sensitive scoring system was used for recall exercises (Part 1–4) because they were identified as more difficult assessments (Laufer et al., 2004). In this regard, approximate answers or partial knowledge would be accepted. Partial knowledge of words would be given credit because of the level of the learners. For instance, in Part 1, the answers '*biernes* (viernes)' and

'*gaspacho* (gazpacho)' would be scored as correct because the pronunciations of /b/-/v/ and /s/-/z/ are similar in Spanish (the italicised answer is the answer provided by learners and the one in brackets is the standard answer). However, '*hueves* (jueves)', '*gitarra* (guitarra)' and '*dereca* (derecha)' would be marked as incorrect despite the high similarity of the word form provided because the pronunciations of /h/-/j/, /gi/-/gui/ and /c/-/ch/ are different in Spanish and learners are expected to know these differences. Slightly misspelled answers in Part 3 and 4 would be accepted because spelling is not a determining factor in these translation assessments but the true connection of form and meaning. For example, in Part 3, '*messa* (mesa)' and '*homber* (hombre)' would be given 0.5 mark and in Part 4, '*perple* (purple)' '*ate* (to eat)' and '*causin* (cousin)' would also be given 0.5 mark considering the intended meaning of the answers provided. A consensus between the researcher, the teacher participant and a native lecturer of Spanish would be sought in case of any discrepancy. Secondly, a strict scoring system was used for the recognition exercises (Part 5–8) because they were identified as less difficult assessments (Laufer et al., 2004). This strict scoring method was reflected in the high degree of similarity and connection of the distractors used in the multiple-choice items as well as the high degree of accuracy required in the answers, that is, only one option was allowed for each test item.

At this juncture, it is important to note that the pre-test and the post-test comprised different test items as the target words involved were rather easy and common at this beginner level. Moreover, it was intended to avoid the "practice effect" (Brown, 1996, p. 79) of taking the same test twice and the potential "testing" (Creswell, 2012, p. 305) effect which was considered as a threat to the internal validity of experimental inferences. Finally, the post-test and the delayed post-test were identical except for the sequence of the items. The items in each section of the delayed post-test were randomly shuffled. The analysis of the vocabulary test content, the vocabulary test format, and a sample of the license certificate for all the graphic source used in this instrument can be found at https://bit.ly/3AJTE0N.

3.2. Validation of instrument

3.2.1. Content validation

The assessment of content validity was based on the three-stage process suggested by Almanasreh et al. (2019). The first development stage was described in Section 3.1. The second judgment-quantification stage involved identifying and inviting experts to validate the content of the vocabulary tests. The third stage was revision and reconstruction of the instrument.

A group of twenty experts were identified via Google Scholar and ResearchGate based on their relevant expertise, research, and publications. Experts were requested to rate each test item on a rating scale of 1 to 4, with 4 being the highest value for the three assessment criteria: Relevance, consistency, and clarity. Relevance indicates how essential is the item to the construct being measured; consistency indicates if the item is logically related to the construct and clarity indicates the adequacy of the syntactic and semantic aspects of the item, that is, whether the item is clearly worded. Moreover, experts were encouraged to provide comments or suggestions for improvement including addition or deletion of any items.

To ensure the quality of this second stage, the test format, the analysis of the test content, the complete word list for the treatments, the description of the research context along with relevant references were provided to adequately inform the experts. From the twenty experts invited, eleven responded and the evaluations collected were analysed quantitatively and qualitatively. These eleven experts consisted of eight native speakers of Spanish: one Argentinian and seven Spanish from Spain (four were working in Spain, two in Malaysia, and one in Poland). Another three experts were non-native speakers of Spanish: two Malaysians from a local university and one Greek from a Greek university.

Content validity index (CVI) was used to quantify the content validity for the vocabulary tests (Polit et al., 2007). The recommended item-level CVI (I-CVI) should not be less than .78 for six or more experts and average scale-level CVI (S-CVI/Ave) should be .90 or higher. The I-CVI for all individual items in both the pre-test and the post-test in terms of the three assessment criteria ranged from .70 to 1.00. Only two items (Pre-test: item 21 and item 33) were found to have an I-CVI less than .78. These two items were subsequently revised and improved based on the qualitative comments. The overall S-CVI/Ave for the 40-item instrument in terms of relevance, consistency, and clarity for both the pre-test and the post-test were higher than .90 (see Table 1). This signified good content validity of the instrument from the quantitative perspective.

The final stage of the content validation involved addressing the experts' qualitative comments and performing revisions. After a critical review of the comments and suggestions, necessary modifications were made to improve the instrument. Following this, an expert evaluation report was prepared and sent to the six experts whose ratings fell between 1 and 2 of the 4-point rating scale for certain test items. For this second-round evaluation, experts were given choices whether to re-evaluate all the affected items or give an overall score on a scale of 1 (unacceptable) to 4 (exceeds expectations) to the modifications made. From these six experts, only four replied with an overall score where two experts rated as 3 (meets expectations) and another two rated as 4 (exceeds expectations). This signified good content validity of the revised instrument from the qualitative perspective.

Table 1. *Content Validity Index (CVI) for the Vocabulary Tests (N = 11)*

Assesment criteria	Pre-test	Post-test
	S-CVI/Ave	*S-CVI/Ave*
Relevance	.97	.98
Consistency	.98	.98
Clarity	.96	.98

Note: N = Number of experts

3.2.2. Pretesting

The validated tests were subjected to pretesting prior to the field study (Wan Mahzan, 2020). Although pretesting is more commonly applied to survey instruments, we are convinced from our observation that conducting pretesting on a small group of Spanish language learners who were similar in kind to the sample of the larger research could increase the validity

of the instrument. Therefore, 15 students were recruited conveniently for each test at the campus library as suggested by Peterson et al. (2017). Two pretesting techniques, cognitive interview including think aloud and cognitive verbal probing and respondent debriefing were used concomitantly.

Students were required to answer two debriefing questions after completing each part of the test and were encouraged to write down their comments in the space indicated. These printed debriefing questions addressed the difficulty and the clarity of the test items. Table 2 shows the results collected from the pretesting. On a scale of 1 to 5, with 1 representing very unclear and 5 representing very clear, respondents reported that they found the instructions and the test items very clear (pre-test: M = 4.82, SD = 0.40; post-test: M = 4.86, SD = 0.35). Similarly, with 1 representing very difficult and 5 representing very easy, respondents reported different levels of difficulty for different parts of the tests. The ratings for all parts of test ranged from 1 to 5 and all SDs exceeded 0.50, this indicated some degrees of variability and diversity among the respondents (pre-test: M = 4.35, SD = 0.78; post-test: M = 4.19, SD = 0.95). Therefore, in terms of difficulty, the respondents generally considered that both tests were easy (M > 4.00). One possible explanation for such a finding may be attributed to all respondents recruited had been studying Spanish for at least one semester.

3.2.3. Item Analysis

With reference to Brown (1996), item analysis for CRTs entails item quality analysis, Difference Index (DI), and *B*-index. As item content analysis is equivalent to the test content validation, the following discussion focuses only on the B-index estimation using the pretesting results. DI estimation could only be done after the field study hence it would be discussed later.

Table 2. *Clarity and Difficulty of the Vocabulary Test Items*

Part of test	Pre-test (n = 13)				Post-test (n = 14)			
	Clarity		*Difficulty*		*Clarity*		*Difficulty*	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Part 1	4.69	0.48	4.23	0.73	4.93	0.27	4.43	0.94
Part 2	5.00	0.00	4.77	0.60	4.93	0.27	4.57	0.65
Part 3	4.85	0.38	4.46	0.66	4.86	0.36	3.86	1.03
Part 4	4.69	0.63	3.69	1.11	4.79	0.43	3.86	1.10
Part 5	4.77	0.60	4.15	0.80	4.79	0.43	3.71	1.07
Part 6	4.77	0.44	4.46	0.78	4.79	0.43	4.14	1.17
Part 7	4.85	0.38	4.31	0.95	4.86	0.36	4.36	0.74
Part 8	4.92	0.28	4.69	0.63	4.93	0.27	4.57	0.94
M	4.82	0.40	4.35	0.78	4.86	0.35	4.19	0.95

Note: N = 27, n = number of valid responses (individuals for pre-test and post-test were different).

B-index indicates the degree to which a CRT item distinguishes the participants who passed from those who failed the test. Put differently, *B*-index compares the item facility (IF) of the participants and to calculate this index, 70% was used as the cut-point for passing the test as recommended by Brown (1996). Table 3 shows that the pre-test could only marginally discriminate between participants who passed the test from those who failed (pre-test: M = 0.23, SD = 0.26) while the post-test could reasonably discriminate between these two groups (post-test: M = 0.35, SD = 0.31). Two possible explanations for such results may be owing to (a) the small number of respondents recruited for the pretesting, and (b) all respondents had some knowledge of Spanish. Moreover, visual inspection on the distribution of the test scores indicated that the data was negatively skewed, an expected feature of CRTs. This made the discrimination between the groups who passed the tests and the ones who failed difficult. Nonetheless, these results would be used for subsequent test quality enhancement along with various other sources of information including the DI and the results of the field study.

3.2.4. Internal consistency reliability

To estimate the reliability of the vocabulary tests, Kuder-Richardson formula 20 (KR-20) was used. According to Brown (1996), KR-20 is the single most accurate estimate that can better avoid the problem of underestimating the internal consistency reliability of a language test if compared with Cronbach's alpha or KR-21. As a result, the KR-20 computed for the pre-test was .883 and the post-test, .932 (see Table 3). These values indicated that both tests were reliable and have a high internal consistency. Besides, the tests had small standard error of measurement (SEM), 2.081 for the pre-test and 1.804 for the post-test. Conceptually, this signified that both tests were precise and consistent in measuring learners' VK.

3.2.5. Equivalence between pre-test and post-test

Compared means test was performed using SPSS version 23 to determine whether the pre-test and the post-test were statistically equivalent. Mann-Whitney *U* test was performed because the Shapiro-Wilk test ($p < .05$) indicated a significant deviation from normality in the pre-test group ($W = .868$, $p = .049$) and the post-test group ($W = .873$, $p = .046$), and the Levene's test ($p < .05$) indicated an equality of variances with $F(1, 25) = .000$, $p = .996$. The *U* test indicated no significant difference in the test scores of the pre-test group (Median = 29, n = 13) and the post-test group (Median = 35, n = 14), $U = 66.000$, $z = -1.217$, $p = .239$, $r = -.23$. Therefore, it can be concluded that the two tests were statistically equivalent.

Table 2. *B-Index for the Pre-test and the Post-test of the Pretesting Groups*

Part of test	Pre-test (n = 13)			Item No.	Post-test (n = 14)		
	IF_{pass}	IF_{fail}	*B-Index*		IF_{pass}	IF_{fail}	*B-Index*
1	1.00	0.80	0.20	1	1.00	0.67	0.33
2	0.75	0.80	-0.05	2	0.82	0.33	0.48
3	0.88	0.20	0.68	3	0.64	0.00	0.64
4	0.88	0.60	0.28	4	1.00	0.00	1.00
5	1.00	0.80	0.20	5	1.00	0.33	0.67
6	1.00	1.00	0.00	6	1.00	1.00	0.00
7	1.00	0.80	0.20	7	1.00	1.00	0.00
8	1.00	1.00	0.00	8	1.00	1.00	0.00
9	1.00	0.80	0.20	9	1.00	1.00	0.00
10	1.00	1.00	0.00	10	1.00	0.67	0.33
11	0.63	0.20	0.43	11	0.82	0.33	0.48
12	1.00	1.00	0.00	12	0.82	0.33	0.48
13	0.63	0.40	0.23	13	1.00	1.00	0.00
14	0.63	0.00	0.63	14	1.00	0.33	0.67
15	0.63	0.00	0.63	15	0.45	0.00	0.45
16	0.88	0.20	0.68	16	1.00	1.00	0.00
17	1.00	0.00	1.00	17	0.73	0.00	0.73
18	0.88	0.40	0.48	18	0.82	0.33	0.48
19	0.63	0.00	0.63	19	0.45	0.00	0.45
20	0.38	0.20	0.18	20	0.91	0.33	0.58
21	1.00	1.00	0.00	21	1.00	1.00	0.00
22	0.88	1.00	-0.13	22	0.82	0.67	0.15
23	0.63	0.40	0.23	23	1.00	0.33	0.67
24	1.00	0.80	0.20	24	0.55	0.00	0.55
25	0.88	0.20	0.68	25	0.64	0.00	0.64
26	0.75	0.80	-0.05	26	0.91	1.00	-0.09
27	0.75	0.80	-0.05	27	1.00	1.00	0.00
28	1.00	1.00	0.00	28	1.00	0.67	0.33
29	1.00	1.00	0.00	29	0.91	0.33	0.58
30	1.00	0.80	0.20	30	1.00	0.67	0.33
31	0.63	0.40	0.23	31	1.00	1.00	0.00
32	1.00	1.00	0.00	32	1.00	1.00	0.00
33	0.25	0.00	0.25	33	0.73	0.00	0.73
34	0.75	0.60	0.15	34	1.00	0.00	1.00
35	0.88	0.80	0.08	35	1.00	0.67	0.33
36	1.00	1.00	0.00	36	1.00	1.00	0.00
37	1.00	0.80	0.20	37	1.00	1.00	0.00
38	0.88	0.60	0.28	38	1.00	0.33	0.67
39	1.00	0.80	0.20	39	1.00	0.67	0.33
40	1.00	0.80	0.20	40	0.91	1.00	-0.09
M	0.85	0.62	0.23	M	0.90	0.55	0.35
SD	0.19	0.35	0.26	SD	0.16	0.40	0.31
KR-20	.883			KR-20	.932		
SEM	2.081			SEM	1.804		

Note: N = Number of valid responses; IF_{pass} = Item Facility for participants who passed the test; IFfail = Item Facility for participants who failed the test; KR-20 = Kuder-Richardson formula 20; SEM = Standard Error of Measurement

3.3. Field study

To complete the steps to assure the psychometric qualities of the CRTs, the Difference Index (DI) which assesses the sensitivity of CRT items to differences in vocabulary instructions (interventions) were computed based on the results of the field study. Table 4 shows the DI of the eight vocabulary test parts for the three groups. Results showed that there was a gradual vocabulary gain for all groups from pre-test to delayed post-test and the test items were sensitive to different vocabulary instructions [i.e., Kruskal-Wallis test, H (2) = 11.609, p = .003 from pre-test to post-test, H (2) = 12.434, p = .002 from pre-test to delayed post-test]. The relatively high DIs for Part 3 to Part 8 indicated good quality of the test items in distinguishing the master group from the non-master group, however, Part 1 and Part 2 required attention.

Based on other indices of item analysis, test items of Part 1 should be retained because the Item Discrimination (ID) indices were 0.40 and above (pre-test: ID = 0.40; post-test: ID = 0.65, see Brown, 1996, p. 70). Nonetheless, this was not the case for Part 2 which showed alarmingly small ID indices (pre-test: ID = 0.10; post-test: ID = 0.05). Therefore, a critical revision was performed for the test items of Part 2 by the researcher, the teacher participant, and a native lecturer of Spanish before they were used in the larger research. As a result, the difficulty of the items was raised by using a sentence input instead of a single word input, that is, the students would be required to identify the target word from a short Spanish sentence. The sentences would be read by the native lecturer of Spanish in a normal reading speed. The distractors would adopt real words which represent common challenges in the early stages of Spanish language learning (see Figure 1).

4. Discussion

The instrument developed is in strict compliance with Read and Chapelle's (2001) framework of vocabulary assessment. Firstly, the instrument adopted a trait definition of vocabulary wherein the test items were discrete, selective and context independent. Although Read and Chapelle advocate for an interactionist approach, it was not practical considering the research objectives, subject, and context. Secondly, as the instrument was intended to provide information about the effects of different interventions on four VK aspects; hence the test purpose could be related to research, or evaluation uses. Thirdly, inferences could be drawn from sub-test level (e.g., multiple scores to inform the gain of individual VK aspects) or whole test (e.g., single score to inform overall VK gain) level. Lastly, the intended effects of the instrument were presumably to promote vocabulary learning among beginner learners of Spanish and provide insights into the potentials of different vocabulary instructions in the mastery of different VK aspects within a FL learning context.

By adhering to the educational measurement theory, we found that the results obtained in this study showed a strong connection with those of Aoiz Pinillos (2022) which indicated high levels of reliability (.82 and .89 for two tests), consistent with previous related studies. This in turn reflects that when an instrument has undergone a principled validation and reliability process, the good psychometric qualities (e.g., adequacy, relevance, and usefulness) would yield reliable estimates of the construct under study, for instance, in this study the test items were sensitive in distinguishing the VK gains of the better and the weaker learners under different interventions (see Table 4).

Table 4. *Difference Index (DI) of the Eight Vocabulary Test Parts*

Test part	Mean score of Item Facility (IF) for			Difference Index for	
	Pre	*Pos*	*Dly*	*Pos-Pre*	*Dly-Pre*
Contro group (n = 10)					
Part 1	0.42	0.44	0.46	0.02	0.04
Part 2	0.92	0.96	0.88	0.04	-0.04
Part 3	0.18	0.86	0.80	0.68	0.62
Part 4	0.00	0.52	0.62	0.52	0.62
Part 5	0.24	0.52	0.60	0.28	0.36
Part 6	0.32	0.84	0.94	0.52	0.62
Part 7	0.20	0.60	0.72	0.40	0.52
Part 8	0.50	0.72	0.76	0.22	0.26
Mean	0.35	0.68	0.72	0.34	0.38
SD	0.28	0.19	0.16	0.24	0.27
Comparison group (n = 12)					
Part 1	0.53	0.55	0.63	0.02	0.10
Part 2	0.93	0.90	0.92	-0.03	-0.02
Part 3	0.28	0.63	0.77	0.35	0.48
Part 4	0.03	0.43	0.55	0.40	0.52
Part 5	0.32	0.53	0.58	0.22	0.27
Part 6	0.38	0.93	0.97	0.55	0.58
Part 7	0.22	0.62	0.82	0.40	0.60
Part 8	0.38	0.77	0.87	0.38	0.48
Mean	0.39	0.67	0.76	0.29	0.38
SD	0.26	0.18	0.16	0.20	0.23
Experimental group (n = 15)					
Part 1	0.64	0.52	0.71	-0.12	0.07
Part 2	0.91	0.96	0.95	0.05	0.04
Part 3	0.11	0.89	0.92	0.79	0.81
Part 4	0.00	0.75	0.88	0.75	0.88
Part 5	0.24	0.72	0.85	0.48	0.61
Part 6	0.35	0.93	0.97	0.59	0.63
Part 7	0.21	0.75	0.77	0.53	0.56
Part 8	0.43	0.89	0.87	0.47	0.44
Mean	0.36	0.80	0.87	0.44	0.51
SD	0.30	0.15	0.09	0.32	0.31

In terms of practical implication, the instrument developed can be used for either research or pedagogical purposes. It is significant not only because it has gone through a rigorous validation process, but also because the entire development process is made explicit for future replications. Moreover, the results obtained from the instrument may render diag-

nostic feedback to students, teachers, and researchers regarding the characteristics of Spanish vocabulary acquisition including the effects of different instructions on different types of VK. For methodological implication, the clarity in methodological rigor is reflected to not only assure the psychometric qualities of the instrument, but also allow proper replications and assessments of potential sources of bias or variability in future studies.

5. Conclusion

This study has addressed the research gap by exemplifying a feasible approach in developing a new instrument for measuring four Spanish VK aspects particularly for research with interventions to counteract the scarcity of valid and reliable Spanish VK tests in the current literature. To present validation evidence systematically, we started by conceptualising and operationalising the construct of VK before designing and developing the instrument rigorously following the four development phases adopted. Then, the instrument went through various types of validation and finally it was tested in a field study. Results showed that the instrument is valid, reliable, and sensitive in distinguishing different groups of learners (master or non-master) under different vocabulary instructions. Additionally, this study has foregrounded evidence to support Laufer and Goldstein's (2004) VK strength hierarchy of four modalities: Active recall (the most difficult), passive recall, active recognition, and passive recognition (the easiest). Therefore, language teachers should be aware that learners' vocabulary performance may vary depending on the types of instructions, assessments and the four modalities of strength. In this regard, more research is encouraged to substantiate this claim.

While this study has provided useful information in relation to the psychometric qualities of a vocabulary measure, there are several limitations and delimitations. The validation process only covered a small portion of the target population. Besides, the non-probability sampling might have introduced participant bias into the study whereby students who participated might have higher interest in the research topic. It is hence recommended that this study should be replicated using a bigger number of samples to substantiate the validity and reliability claims. Moreover, more aspects of validity could be assessed since validation is a continuing process. For instance, criterion validity could be evaluated through examining the correlations between the vocabulary performance and other language related assessments like writing, speaking, reading, and listening. Likewise, Rasch measurement is also encouraged when an adequate sample size is available. To end, we encourage language teachers and researchers to become vocabulary assessment literate and this study has provided a solid example that is consistent with current educational measurement theory.

6. Acknowledgements

We are grateful for the support from Universiti Kuala Lumpur. We would also like to extend our appreciation to Alejandra Ma Rodríguez Torres, Apolinar García Paredes, and Dra. Inmaculada Clotilde Santos Díaz for their relentless efforts in revising the Spanish abstract. Lastly, our heartfelt thanks to the anonymous reviewers and editors whose insightful feedback played a vital role in enhancing the overall excellence of this article.

7. References

Almanasreh, E., Moles, R., & Chen, T. F. (2019). Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy*, *15*(2), 214–221. https://doi.org/10.1016/j.sapharm.2018.03.066

Aoiz Pinillos, M. (2022). Creation and validation of a bilingual test to estimate aural and written vocabulary size. *Porta Linguarum Revista Interuniversitaria de Didáctica de Las Lenguas Extranjeras*, *38*, 247–263. https://doi.org/10.30827/portalin.vi38.23606

Brown, J. D. (1996). *Testing in language programs*. Prentice Hall Regents.

Creswell, J. W. (2012). *Educational research: Planning, conducting and evaluating quantitative and qualitative research* (4th ed.). Pearson Education Inc.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.

Davies, M. (2006). *A frequency dictionary of Spanish: Core vocabulary for learners*. Routledge.

Kremmel, B. (2018). *Development and initial validation of a diagnostic computer-adaptive profiler of vocabulary knowledge* [PhD Thesis]. University of Nottingham.

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, *21*(2), 202–226. https://doi.org/10.1191/0265532204lt277oa

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

Ma, Q., & Lee, H. Y. (2019). Measuring the vocabulary knowledge of Hong Kong primary school second language learners through word associations: Implications for reading literacy. In B. L. Reynolds & M. F. Teng (Eds.), *English Literacy Instruction for Chinese Speakers* (pp. 35–56). Palgrave Macmillan. https://doi.org/10.1007/978-981-13-6653-6

Nakata, T., & Webb, S. (2016). Does studying vocabulary in smaller sets increase learning? The effects of part and whole learning on second language vocabulary acquisition. *Studies in Second Language Acquisition*, *38*(3), 523–552. https://doi.org/10.1017/S0272263115000236

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

Nation, I. S. P. (2020). The different aspects of vocabulary knowledge. In S. A. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (1st ed., pp. 15–29). Routledge. https://doi.org/10.4324/9780429291586.2

Nation, I. S. P., & Hunston, S. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139858656

Onwuegbuzie, A. J., & Collins, K. M. T. (2007). A typology of mixed methods sampling designs in social science research. *The Qualitative Report*, *12*(2), 281–316.

Peterson, C. H., Peterson, N. A., & Powell, K. G. (2017). Development cognitive interviewing for item development: Validity evidence based on content and response processes. *Measurement and Evaluation in Counseling and Development*, *50*(4), 217–223. https://doi.org/10.1080/07481756.2017.1339564

Polit, D. F., Beck, C. T., & Owen, S. v. (2007). Focus on research methods: Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing and Health*, *30*(4), 459–467. https://doi.org/10.1002/nur.20199

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

Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing*, *18*(1), 1–32. https://doi.org/10.1177/026553220101800101

Sánchez Gutiérrez, C. H. (2021). Morphology and language teaching. In A. Fábregas, V. Acedo-Matellán, G. Armstrong, M. C. Cuervo, & I. Pujol Payet (Eds.), *The Routledge Handbook of Spanish Morphology* (1st ed.). Routledge. https://doi.org/10.4324/9780429318191.42

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

Sinclair, J., & Renouf, A. (1988). A lexical syllabus for language learning. In R. Carter & M. Mccarthy (Eds.), *Vocabulary and Language Teaching* (pp. 140–158). Routledge.

Verspoor, M., & Lowie, W. (2003). Making sense of polysemous words. *Language Learning*, *53*(3), 547–586.

Wan Mahzan, M., Alias, N., & Ismail, I. (2020). Unboxing the design of English as a second language (ESL) learning video game for indigenous learners: An empathic design-based approach. *Asia Pacific Journal of Educators and Education, 35*(2), 39–56. https://doi.org/10.21315/apjee2020.35.2.3

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, *27*, 33–52. https://doi.org/10.1017/S0272263105050023

Webb, S. (2007). Learning word pairs and glossed sentences: The effects of a single context on vocabulary knowledge. *Language Teaching Research*, *11*(1), 63–81. https://doi.org/10.1177/1362168806072463