

**AUTHOR ACCEPTED MANUSCRIPT**

**NOTE: This article is under copyright and the publisher should be contacted for permission to re-use or reprint the material in any form. Please cite the final published version.**

Daidone, D. (2023). The relationship between self-rated vocabulary knowledge and accuracy of phonological forms. In I. Checa-García & L. Marqués-Pascual (Eds.), *Current perspectives on Spanish lexical development* (pp. 49-72). de Gruyter Mouton.  
<https://doi.org/10.1515/9783110730418-003>

# The relationship between self-rated vocabulary knowledge and accuracy of phonological forms

Danielle Daidone

University of North Carolina Wilmington

## 1 Introduction

For literate learners of a language, an entry in their second language (L2) mental lexicon for a word includes three core aspects: an orthographic representation, a semantic representation, and a phonological representation (see Ramus et al. 2010). Thus, an important component of lexical knowledge is knowing the phonological form of words, but research on L2 vocabulary acquisition has typically neglected to examine phonological forms in favor of focusing on orthographic and semantic representations. This is reflected in Beglar and Nation's (2014) review of vocabulary assessment, which includes almost entirely written tests of vocabulary. Well-known tests of English vocabulary knowledge include Nation's Vocabulary Levels Test and its updated versions (Schmitt, Schmitt and Clapham 2001), as well as the Vocabulary Size Test (Nation and Beglar 2007). For these types of vocabulary tests, the test taker matches written words with their definitions. While the Vocabulary Knowledge Scale (Wesche and Paribakht 1996) mixes both receptive and productive knowledge, it also only focuses on the written form, testing learners on a scale to assess the depth of their knowledge, from "I don't remember having seen this word before" to "I can write this word in a sentence: \_\_\_\_\_ (Write a sentence.)". Finally, typical "Yes/No" tests present participants with written words and nonwords, asking them to indicate which words they know. As Beglar and Nation (2014) state, "despite the importance of developing measures of aural vocabulary size, very little empirical work on tests designed to measure this construct exists" (p. 4).

Milton (2009) argues that this bias toward concentrating on orthographic forms is likely due to the ease of written tests and the unstated assumption in the field that knowledge of a written form implies knowledge of its phonological form (p. 93). However, research in L2 phonology has shown that L2 learners' phonological representations are often inaccurate or less detailed than those of native speakers (e.g., Cook et al. 2016; Cutler, Weber and Otake 2006; Daidone and Darcy 2014; Darcy, Daidone and Kojima 2013; Darcy and Thomas 2019; Llompart 2021; Melnik and Peperkamp 2019). While some researchers have taken for granted that learners know the frequent words used in their study (Cutler, Weber and Otake 2006; Llompart 2021; Melnik and Peperkamp 2019) and others have manipulated frequency as a variable (Cook et al. 2016), the presence of inaccurate representations appears to be true even when an effort is made to ensure that learners know the words that were tested.

Darcy, Daidone, and Kojima (2013) tested intermediate and advanced English-speaking learners of German and Japanese on their perception and phonological lexical representations of difficult contrasts in these languages, specifically front rounded vowels for German and consonant length in Japanese. They found that the learners of German were able to discriminate front and back rounded vowels, and the learners of Japanese were able to discriminate singleton and geminate consonants. However, a lexical decision task revealed that both intermediate and

advanced learners of Japanese had difficulty rejecting nonwords with a singleton consonant if the real word contained a geminate consonant; for example, they accepted \**kipu* [kipu] as a word when the real word is *kippu* /kippu/ ‘ticket.’ Similarly, the intermediate German learners had difficulty rejecting nonwords with a back rounded vowel if the real word contained a front rounded vowel. This is despite the fact that the words for the Japanese experiment were common words chosen from the first-year and second-year textbooks used by the students, and the participants in the German experiment were tested on their familiarity with the stimuli, and any participant with low written familiarity with the words was excluded. Darcy and Thomas (2019) also assessed familiarity for the words used in their experiment that examined the encoding of consonant clusters in L2 lexical representations by Korean learners of English. They found that participants rated all of the words as “very familiar” when given three options for familiarity, but nevertheless they tended to falsely accept test nonwords with an epenthetic vowel such as *b[ɔ]lue* for *blue*.

Similar findings on inaccurate lexical representations have also been reported for Spanish. Daidone and Darcy (2014) tested intermediate and advanced English-speaking learners on their lexical representations containing Spanish /r/, /r/, and /d/ and their perception of these phonemes. They found that despite high accuracy in a discrimination task, in which the advanced learners’ results did not differ from those of native speakers, in a lexical decision task both intermediate and advanced learners had difficulty rejecting nonwords with the incorrect rhotic, such as \**quierro* [kjero] for the real word *quiero* /kjero/ ‘I want’. Learners had difficulty with these experimental words even though the researchers mainly used stimuli from the *Beginning Spanish Lexicon*, which provides data about words that are included in beginner Spanish textbooks (Vitevitch, Stamer and Kieweg 2012), and they checked that no participants had generally low familiarity with the words used.

Given these findings on learners’ inaccurate phonological representations, it is not clear that knowledge of the meaning of a written form necessarily entails an accurate phonological form. Moreover, studies specifically examining the effect of orthography on lexical encoding accuracy have found mixed results. Some researchers have shown that orthography can help learners establish a difference between new non-native phonological contrasts (Escudero, Hayes-Harb and Mitterer 2008; Escudero, Simon and Mulak 2014; Showalter and Hayes-Harb 2013). For example, Escudero, Hayes-Harb, and Mitterer (2008) found that those Dutch-speaking learners of English exposed to orthographic forms of new words were able to establish a contrast between words differing in the difficult contrast /æ-ɛ/, whereas those participants only exposed to auditory input were less accurate. On the other hand, researchers have also found a negative effect of exposing learners to orthography when building new lexical representations, especially when the orthography does not match the sound-spelling correspondences of their first language (Escudero, Simon and Mulak 2014; Hayes-Harb, Brown and Smith 2018; Hayes-Harb and Cheng 2016; Hayes-Harb, Nicol and Barker 2010; Mathieu 2016; Rafat 2016). Escudero, Simon, and Mulak (2014) revealed that Spanish learners of Dutch were aided by the presence of orthographic forms only if the sound-spelling correspondences were similar to those in Spanish; new sound-spelling correspondences in Dutch hindered their performance. Similarly, Rafat (2016) showed that English-speaking learners of Spanish were negatively influenced by seeing Spanish orthographic forms that differed from English sound-spelling correspondences; for instance, the presence of <z> (/s/ or /θ/ in Spanish, depending on the dialect) often resulted in the pronunciation [z] as would be

found in English. Other studies have observed no effect of exposing learners to orthography when acquiring words (Durham et al. 2016; Hayes-Harb and Hacking 2015; Showalter and Hayes-Harb 2015; Simon, Chambless and Kickhöfel Alves 2010). For example, Simon and colleagues taught English-speaking participants French novel words containing /u/ and /y/, either with orthography or without. They reported no difference in the accuracy of these vowels in word learning or perception between the group that was given orthography and the group that was exposed only to auditory forms. In sum, these studies show that orthography may help, but just as often may have no influence or even hinder learners' phonological acquisition of words. At the very least, this research reveals that learning the orthographic form of a word does not guarantee its phonological form will be remembered correctly.

This gap between knowledge of orthographic and phonological forms can have real-world consequences for learners, who may know a word in its written form but are unable to recognize it in speech, as evidenced in English by Carney (2021). He found that Japanese-speaking learners of English had difficulty with listening comprehension even for texts that consisted of high-frequency vocabulary and orthographically known words. Therefore, it is unsurprising that written vocabulary knowledge is not a strong predictor of listening comprehension in English, while aural vocabulary knowledge very strongly predicts listening comprehension (Masrai 2020).

Given this disconnect between orthographic knowledge and phonological knowledge, researchers in English language teaching have begun to see the value of creating tests for learners' knowledge of the phonological form of words, as seen by the development of the Listening Vocabulary Levels Test (McLean, Kramer and Beglar 2015). For measures of aural vocabulary size, researchers have also used dictations, in which participants write down aurally presented sentences, and the number of corrected written keywords is tallied, or an aural "Yes/No" test such as AuralLex (Milton and Hopkins 2006), in which learners hear words and nonwords and indicate whether they know them or not. Some existing tests of productive knowledge do assess oral production, such as the Boston Naming Test (Kaplan, Goodglass and Weintraub 2001) and the Peabody Picture Vocabulary Test (PPVT) (Dunn and Dunn 1997). However, these tests were created for use with children or adult native speaker populations and have haphazardly been adapted for use with adult L2 learners.

For Spanish, few standardized tests of vocabulary knowledge exist for adult L2 learners, and they are exclusively written. The Lextale-Esp (Izura, Cuetos and Brysbaert 2014) asks learners to indicate which words they know from a list, and X\_Lex (Meara 2005) asks learners to indicate whether each individually presented written form is a real word. Standardized aural tests of vocabulary have yet to be developed for Spanish, but given the transparent orthography of Spanish, it could be argued that written tests of vocabulary knowledge may in fact be sufficient for L2 learners. In the absence of such aural tests of L2 Spanish vocabulary, especially ones that test phonological form together with meaning, it is not clear whether previous results that have pointed to inaccurate phonological representations for Spanish words are truly due to a disconnect between learners' knowledge of the orthographic form of a word and its phonological accuracy, or are related more to experimental design decisions made in previous studies.

In Darcy and Daidone (2014), as well as other studies, the researchers did not take vocabulary knowledge into account on a trial-by-trial basis, and thus unknown words may have depressed accuracy levels in their results. They also did not look at differences in vocabulary

knowledge as anything more than an exclusion criterion, when in fact gradation in vocabulary knowledge could help explain differences in performance between participants. Furthermore, that study, along with the majority of other studies examining L2 lexical representations, used an auditory lexical decision task in which participants heard stimuli and had to decide if they were real words or not. It may be that L2 learners know the correct pronunciation of a word, but they are willing to consciously accept slight deviations of the type tested in the experimental conditions because they are less confident in their answers or are used to hearing other non-native speakers, and thus have a lower threshold for acceptance. Therefore, a more in-depth analysis of participants' knowledge of the words used in an experiment may reveal that learners of Spanish do in fact display accurate phonological representations in accordance with their knowledge of the semantic and orthographic representations of those words, particularly when assessed with a task that requires them to choose the correct form of a word rather than indicate whether a form is a real word. The current study examines whether this is indeed the case.

## 2 Method

### 2.1 Research Question

Does self-reported knowledge of the meaning of Spanish written words correspond with accuracy in their phonological form as evidenced by a forced choice task? Does this differ by proficiency level?

In order to address this question, the current study compares intermediate and advanced L2 learners' abilities to choose the correct phonological form of a word in a forced choice task with their self-rated familiarity with words as measured with a vocabulary rating task. Participants also completed a language background questionnaire to collect demographic information and to determine whether the learners understood sound-spelling correspondences in Spanish. These tasks are described in detail in the following sections.

### 2.2 Forced Choice Task

The forced choice task tested the Spanish /r-d/ ("tap-d"), /r-r/ ("tap-trill"), /r-d/ ("trill-d"), and /f-p/ contrasts. The first three contrasts served as test contrasts, and the last served as a control. Whereas the /f-p/ contrast exists in English, the /tap-d/, /tap-trill/ and /trill-d/ contrasts represent new sounds for English-speaking learners in various ways.

Although English has a single rhotic while Spanish has two – the tap and trill – the /tap-trill/ contrast has been shown to be discriminable for learners at all levels and even naïve English listeners (Daidone and Darcy 2014, 2021; Herd 2011; Rose 2010). Nevertheless, English-speaking learners perceptually assimilate both the tap and the trill chiefly to English /ɹ/ (Rose 2012), which may explain why this contrast is difficult for learners to encode accurately in words (Daidone and Darcy 2014, 2021).

While [r] exists in English, at least for North American speakers, it is an allophone of /t/ and /d/ rather than a separate phoneme (Ladefoged and Johnston 2011: 74). Thus, English-

speaking learners must acquire the Spanish tap as a separate phoneme rather than as an allophone of /d/. The different representational status of these sounds in English and Spanish may help explain why the /tap-d/ contrast has been found to be more difficult than /tap-trill/ in perception (Daidone and Darcy 2014, 2021; Herd 2011; Rose 2010), since sounds which are allophones in a speakers' first language are perceived to be perceptually similar (Boomershine et al. 2008). However, /tap-d/ has also been found to be more accurate in lexical encoding than /tap-trill/ (Daidone and Darcy 2014, 2021).

Despite the trill being a new rhotic for English-speaking learners, the /trill-d/ contrast has been shown to be relatively easy for learners, perhaps because the trill and /d/ are mainly perceptually assimilated to different English sounds (Rose 2012). Studies have found the /trill-d/ contrast to be more accurate than /tap-trill/ or /tap-d/ in perception and lexical encoding (Daidone and Darcy 2014, 2021; Herd 2011; Rose 2010).

Finally, the phonemes /f/ and /p/ exist in both English and Spanish, although /p/ has different phonetic properties across the two languages (Hualde 2005: 150). The contrast /f-p/ has served as a control contrast in previous research and has been shown to be more accurate in perception and lexical encoding than /tap-d/, /tap-trill/ and /trill-d/ (Daidone and Darcy 2014, 2021).

In sum, these contrasts were chosen because much is known about their L2 discriminability and lexical encoding accuracy from previous investigations (Daidone and Darcy 2014, 2021; Herd 2011; Herd, Sereno and Jongman 2015; Rose 2010). Nevertheless, the relationship between lexical encoding accuracy and gradations in vocabulary knowledge for words with these sounds has not been examined. Furthermore, while the orthographic representations of these sounds are transparent in intervocalic position, the Spanish orthographic contrast between the tap and trill, <r> vs. <rr> (e.g. *pero* /'pe.ro/ 'but' vs. *perro* /'pe.ro/ 'dog'), is new for English-speaking learners, for whom both a single and a double <r> represent /ɾ/ in their first language (e.g. *carrot* and *carat* /'kæ.ɹət/). Therefore, orthographic knowledge can also be examined as a factor in learners' lexical encoding accuracy.

In the forced choice task, participants heard two auditory stimuli containing the sounds of one of the aforementioned contrasts and decided which was the real Spanish word, such as *quiero* [k̄j̄ero] 'I want' vs. the nonword \**quierro* [k̄j̄ero] (see Table 1 for examples).<sup>1</sup>

Table 1: Example stimuli for the forced choice task

		Word		Nonword	
		Orthography	IPA	Orthography	IPA
/tap-trill/	/r-*r/	aburrido	/a.bu.'ri.do/	aburido	/a.bu.'ri.do/
		tierra	/'t̄je.ra/	tiera	/'t̄je.ra/
(test)	/r-*r/	dinero	/di.'ne.ro/	dinerro	/di.'ne.ro/
		parece	/pa.'re.se/	parrece	/pa.'re.se/
/tap-d/	/r-*d/	cultura	/kul.'tu.ra/	cultuda	/kul.'tu.da/

<sup>1</sup>Full stimuli list, data, and analyses are available at <https://osf.io/4venh/>

(test)		fuera	/ˈfue.ra/	fueda	/ˈfue.da/
	/d-*/r/	miedo	/ˈmie.do/	miero	/ˈmie.ro/
		médico	/ˈme.di.ko/	mérico	/ˈme.ri.ko/
/trill-d/	/r-*/d/	ocurre	/o.ˈku.re/	ocude	/o.ˈk.ude/
		arregla	/a.ˈre.gla/	adegla	/a.ˈde.gla/
(test)	/d-*/r/	estado	/es.ˈta.do/	estarro	/es.ˈta.ro/
		todavía	/to.da.ˈbi.a/	toravía	/to.ra.ˈbi.a/
/f-p/	/f-*/p/	jefe	/ˈxe.fe/	jepe	/ˈxe.pe/
		oficina	/o.fi.ˈsi.na/	opicina	/o.pi.ˈsi.na/
(control)	/p-*/f/	grupo	/ˈgru.po/	grufo	/ˈgru.fo/
		zapato	/sa.ˈpa.to/	zafato	/sa.ˈfa.to/

Words ranged between 2 and 4 syllables, with the target phoneme appearing in intervocalic position as the onset of the 2<sup>nd</sup>, 3<sup>rd</sup>, or 4<sup>th</sup> syllable. The stimuli were recorded in a sound booth by two native Spanish speakers: 1) a male speaker from Costa Rica and 2) a female speaker from Puerto Rico. The speakers produced the stimuli with a standard Spanish pronunciation. Therefore, all /r/ tokens were realized with one occlusion, all /r/ tokens were realized with at least two occlusions, and /d/ was realized as an approximant [ð].

There were 20 words for each of the contrasts, for example, 10 real words containing tap and 10 real words containing trill for the /tap-trill/ contrast. Each word-nonword pair was presented twice, once with the male speaker producing the real word and once with the female speaker producing the real word, theoretically resulting in a total of 40 trials per contrast. However, due to a recording error, the nonword *genedal* (cf. *general* /xeneral/ ‘general’) was not available for the female speaker, resulting in one less trial for the /tap-d/ contrast, for a total of 39 trials. Conversely, a coding error resulted in the duplication of the *correcto-correcto* (nonword-word) trial, bringing the total number of trials for the /tap-trill/ contrast to 41. Thus, there were 160 trials in the task, but with one less trial for /tap-d/ and one additional trial for /tap-trill/. During each trial in the task, a fixation cross appeared on the screen while participants listened to a stimulus spoken by the male speaker, followed by a 500 ms pause, and then a stimulus spoken by the female speaker, always in that order. Participants had 5000 ms from the beginning of the trial to respond, and the time between trials was 1000 ms. Participants also had to complete 10 practice trials at the beginning of the task. These trials, spoken by one female native Spanish speaker from Colombia, contained 5 words and 5 corresponding nonwords that differed in sounds other than the contrasts used in the test and control trials (e.g. *nada* ‘nothing’ vs. *bada*; *duda* ‘doubt’ vs. *dida*). The 5 word-nonword pairs occurred twice, once with the word first and once with the nonword first. For each practice trial, participants received feedback on whether their answer was correct, incorrect, or too slow. To proceed to the rest of the task, participants needed to score at least 8 out of 10, or the practice phase was repeated. Participants completed the forced choice task through a web browser with jsPsych (de Leeuw 2015) in about 10 minutes, with one break in the middle of the task. Trials were divided so that each block contained roughly an equal number per contrast, and trials were randomized within each block.

### 2.3 Vocabulary Task

In order to examine a range of self-reported vocabulary knowledge rather than simply known/unknown, the vocabulary task for the current study was adapted from the Vocabulary Knowledge Scale (Wesche and Paribakht 1996). Given that the majority of studies on L2 vocabulary size as well as word familiarity checks for other types of L2 studies use Yes/No tests of receptive vocabulary knowledge only, the scale of self-reported knowledge used in the current study did not require production, in contrast to the original Vocabulary Knowledge Scale. To clarify, while this self-assessment asked about productive knowledge, it did not require learners to write translations for the words or write them in a sentence. The 6-point scale given to learners was as follows, with only the text of each point on the scale visible rather than the numbers:

- 1) I didn't know this was a word
- 2) I recognize this word but I don't know what it means
- 3) I recognize this word and have a vague idea of what it means
- 4) I recognize this word and know more or less what it means
- 5) I know this word and can provide a translation in English
- 6) I know this word well, can provide a translation in English, and can use this word while speaking Spanish

#### 2.4 Language Background Questionnaire

Each participant completed a language background questionnaire to elicit demographic information such as their age, gender, level of education, and history of residence, as well as their language learning history. In order to check the L2 learners' knowledge of Spanish sound-spelling correspondences, this questionnaire also asked them to "describe the difference in pronunciation of 'r' as in *pero* and 'rr' as in *perro*. If you don't know, please indicate that."

#### 2.5 Participants

Participants in this study were intermediate and advanced English-speaking learners of Spanish, who constituted the experimental groups, and Spanish-speaking learners of English, who served as a control group. The intermediate learners were undergraduate Spanish majors and minors enrolled in a Spanish course at the fifth-semester level or higher. The advanced learners were graduate students who had taken graduate courses in Spanish. A postdoc in Hispanic linguistics also served as an advanced participant. Most of the advanced learners were teaching Spanish and studying Hispanic linguistics or Hispanic literatures and cultures, as were most of the native Spanish speakers. The English-speaking learners all grew up in monolingual households in which only English was spoken. In total, 42 L2 learners of Spanish and 11 native speakers were tested. However, 6 participants were excluded from the analyses for various reasons, such as failing the hearing screening (described below) or having too many timeouts on the forced choice task. This left data from 26 intermediate learners, 12 advanced learners, and 8 native Spanish speakers for the analyses. Demographic information about the participants is presented in Table 2.

Table 2: Demographic information for the participants



	Intermediate L2 Learners <i>N</i> =26	Advanced L2 Learners <i>N</i> =12	Native Spanish Speakers <i>N</i> =8
Age at testing (years)	20.7 (2.5)	26.3 (3.3)	29.5 (2.5)
Age of onset for L2 learning <sup>a</sup>	12.7 (2.7)	13.8 (2.0)	10.6 (8.2)
Residence in a Spanish-speaking country (months)	1.0 (2.3)	13.1 (11.8)	
Age of arrival in the US			24.9 (4.0)
Self-rated L2 speaking ability (0-6)	3.1 (1.4)	5.6 (0.7)	5.1 (1.2)
Self-rated L2 listening ability (0-6)	3.5 (1.4)	5.4 (0.7)	5.6 (0.5)
Self-rated L2 reading ability (0-6)	4.0 (1.1)	5.7 (0.9)	5.4 (1.1)
Self-rated L2 writing ability (0-6)	3.9 (1.5)	5.5 (0.9)	5.1 (1.1)
Gender	20 female	6 female	3 female, 1 non- binary

*Note.* “L2” in the variables refers to Spanish for the English-speaking learners and English for the Spanish speakers. Means are given for rows 1-8, with standard deviations in parentheses.

<sup>a</sup>One Spanish-speaking participant listed their age of onset for L2 learning as “Middle school but formal instruction at the age of 18.” This was not included in the summary statistics in the table.

## 2.6 Procedure

Participants completed the forced choice task, vocabulary knowledge task, and language background questionnaire as part of a larger study. After consenting to do the study, participants completed a bilateral hearing screening with 1000 Hz, 2000 Hz, and 4000 Hz pure tones at 20 dB HL, following the recommendations of Reilly, Troiani, Grossman, and Wingfield (2007). Pulsed tones were presented randomly, one time for each ear, and participants needed to indicate that they heard the tone by pressing the space bar. If an individual missed a tone, all of the tones were repeated once more before the test indicated a failure to pass. All participants needed to pass the hearing screening with 100% accuracy in order to proceed with the study. They were given a maximum of three attempts to pass the hearing screening, if necessary, after attempting to reduce any external noise that could be interfering. This task was administered through jsPsych and took approximately two minutes. Participants next completed a lexical decision task and oddity task that are not discussed in the current study, followed by the forced choice task. They then did a series of cognitive tasks and a vocabulary size task that are not analyzed here. Lastly, they completed the language background questionnaire and the vocabulary knowledge task. All testing took place one-on-one with the researcher in person. Participants wore Sennheiser HD 515 over-ear headphones for the tasks that presented auditory stimuli. The entire experiment lasted approximately 65-75 minutes and each person was paid \$15 for their participation.

### 3 Results

In order to analyze the vocabulary knowledge task, learners' vocabulary ratings were converted to numerical values, with "1" representing no knowledge of the word and "6" representing high familiarity. Because equal difference could not be assumed between each value, vocabulary rating was treated as an ordinal variable. For example, the difference between "I didn't know this was a word" and "I recognize this word but I don't know what it means" is presumably conceptually larger than between "I know this word and can provide a translation in English" and "I know this word well, can provide a translation in English, and can use this word while speaking Spanish". Table 3 displays the median rating and range of ratings in parentheses for each group for each of the experimental contrasts.

Table 3: L2 learners' median vocabulary knowledge rating by contrast

	/tap-trill/	/tap-d/	/trill-d/	/f-p/
Intermediate	6 (1-6)	6 (1-6)	6 (1-6)	6 (1-6)
Advanced	6 (3-6)	6 (all 6)	6 (5-6)	6 (5-6)

#### 3.1 Relationship between vocabulary rating and forced choice accuracy

If self-rated vocabulary knowledge corresponds to the phonological accuracy of a word, we would expect a correlation between the rating given to a word and its accuracy in the forced choice task. In order to test this, Kendall tau-b correlations were run in R for intermediate and advanced learners individually, following the advice of Khamis (2008) for analyzing ordinal variables with a small number of levels. In these correlations, the variables were accuracy in a trial (0 or 1) and vocabulary rating for the word in that trial (1, 2, 3, 4, 5, 6). For the intermediate learners, there was a weak, positive correlation between accuracy in a trial and the vocabulary rating given to that word,  $\tau_b = .12, p < .001$ , with a 95% confidence interval of .08 to .15 for  $\tau_b$ . For the advanced learners, although the initial analysis found a very weak but significant correlation,  $\tau_b = .09, p < .001$ , the 95% confidence interval went through 0 ([-.02, .20]), suggesting that there is no relationship between these variables. These findings of a weak to no correlation between vocabulary rating and accuracy in the forced choice task may be because the words were generally well known to the participants, especially the advanced learners, as seen in Table 3.

Additional analyses were conducted in order to determine if vocabulary rating could predict accuracy in the forced choice task for each contrast. Because the vocabulary ratings did not display much variation for the advanced group, both L2 groups were combined for these analyses. Figure 1 displays the average accuracy for each contrast in the forced choice task for each point along the vocabulary scale, with both L2 groups included. Diamonds represent mean values and the lines at 50% indicate chance level performance.

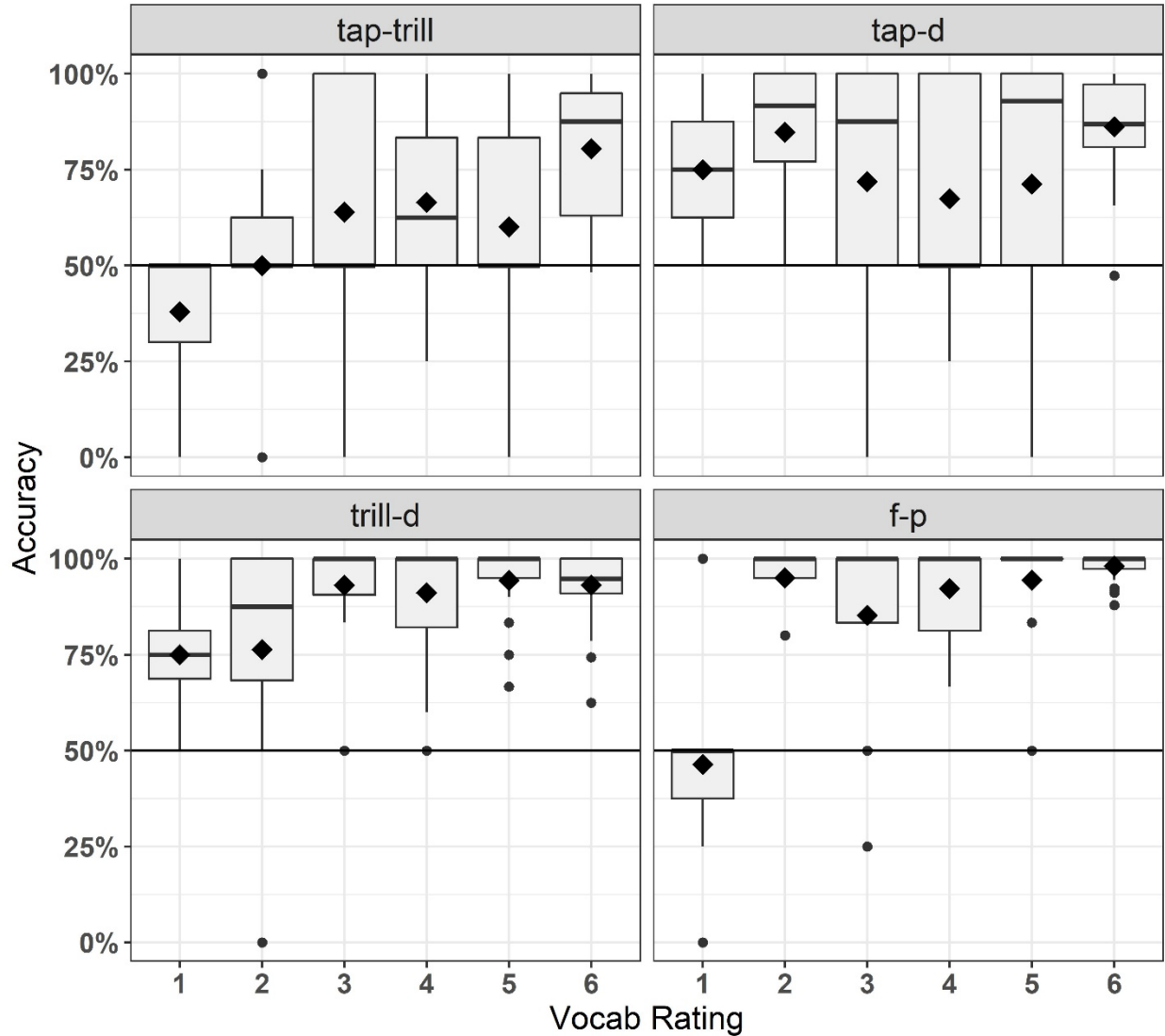


Figure 1: Accuracy in the forced choice task by vocabulary rating and contrast

For each contrast (*/tap-trill/*, */tap-d/*, */trill-d/*, and */f-p/*), a logistic regression was run with accuracy in each trial as the dependent variable (0 or 1) and orthogonal polynomials of the numeric vocabulary rating for that trial's word as the independent variables (vocab score, vocab score squared, vocab score cubed, vocab score to the fourth, and vocab score to the fifth). By analyzing vocab score in this way, it was possible to retain the ordered nature of the scale, that is, treat vocabulary rating as an ordinal variable, while at the same time assessing whether the relationship between vocabulary score and accuracy in the forced choice task was linearly increasing, or whether adding a curve to the regression line was needed in order to best model the data. For example, if only vocab score were significant, then the best fit for the data would be a straight line (i.e. a linear equation), whereas if vocab score squared were also significant, this would mean that adding a U-shaped curve (i.e. a quadratic equation) to the regression line would better model the data. As displayed in Table 4, only the linear trend was significant for the */tap-trill/* contrast, and the model was significant compared to the empty model with the intercept only,  $\chi^2(5) = 57.27, p$

< .001. Thus, there was a linear relationship between vocab score and accuracy for the /tap-trill/ contrast in the forced choice task, such that a higher vocab score predicted higher accuracy.

Table 4: Summary of logistic regression analysis for /tap-trill/ contrast

Predictor	B	B 95% CI	Std Error B	z-value	<i>p</i>	
(Intercept)	1.22	[1.10, 1.33]	0.059	20.477	<.001	***
Vocab	15.40	[11.31, 19.56]	2.099	7.339	<.001	***
Vocab <sup>2</sup>	1.82	[-2.37, 5.88]	2.098	0.865	0.387	
Vocab <sup>3</sup>	2.87	[-1.15, 6.88]	2.044	1.403	0.160	
Vocab <sup>4</sup>	2.25	[-1.79, 6.30]	2.061	1.091	0.275	
Vocab <sup>5</sup>	2.13	[-1.88, 6.14]	2.038	1.044	0.296	

Note.  $\chi^2(5) = 57.27, p < .001$ . B = unstandardized regression coefficient. Numbers in brackets indicate the lower and upper limits of a 95% confidence interval. \*  $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$

Table 5 displays the results of the logistic regression for the /tap-d/ contrast, which was significant,  $\chi^2(5) = 26.55, p < .001$ . For this contrast, both the linear and the quadratic trends were significant. Therefore, a linearly increasing model was not sufficient to explain the data, and a higher vocab score did not always predict higher accuracy for the /tap-d/ contrast. Instead, adding a U-shaped curve (the quadratic equation) to the regression resulted in a better fit for the data. This trend is reflected in the averages in Figure 1 as well, where we see that the average accuracy by vocabulary rating for this contrast dips for ratings 3 through 5.

Table 5: Summary of logistic regression analysis for /tap-d/ contrast

Predictor	B	B 95% CI	Std Error B	z-value	<i>p</i>	
(Intercept)	1.59	[1.46, 1.72]	0.068	23.460	<.001	***
Vocab	9.77	[5.02, 14.23]	2.330	4.194	<.001	***
Vocab <sup>2</sup>	7.00	[2.53, 12.05]	2.353	2.973	0.003	**
Vocab <sup>3</sup>	1.18	[-4.20, 5.63]	2.396	0.494	0.621	
Vocab <sup>4</sup>	0.90	[-3.57, 5.61]	2.299	0.391	0.696	
Vocab <sup>5</sup>	-0.03	[-4.45, 4.20]	2.190	-0.012	0.990	

Note.  $\chi^2(5) = 26.55, p < .001$ . B = unstandardized regression coefficient. Numbers in brackets indicate the lower and upper limits of a 95% confidence interval. \*  $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$

The results of the logistic regression for the /trill-d/ contrast are shown in Table 6. This model was significant,  $\chi^2(5) = 26.55, p < .001$ . For this contrast, only the linear relationship was significant, showing that an increase in vocabulary rating predicted an increase in accuracy in the forced choice task.

Table 6: Summary of logistic regression analysis for /trill-d/ contrast

Predictor	B	B		Std Error B	z-value	p
		95% CI				
(Intercept)	2.29	[2.12, 2.46]		0.087	26.455	<.001
Vocab	9.04	[3.08, 14.65]		2.938	3.078	0.002
Vocab^2	-2.04	[-7.86, 3.53]		2.887	-0.707	0.479
Vocab^3	2.88	[-2.77, 8.47]		2.836	1.017	0.309
Vocab^4	3.92	[-1.79, 9.53]		2.870	1.367	0.172
Vocab^5	-2.84	[-10.06, 3.45]		3.374	-0.842	0.400

Note.  $\chi^2(5) = 15.07, p = .010$ . B = unstandardized regression coefficient. Numbers in brackets indicate the lower and upper limits of a 95% confidence interval. \*  $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$

Table 7 displays the results of the logistic regression for the /f-p/ contrast, which was significant,  $\chi^2(5) = 63.14, p < .001$ . Similar to the /tap-trill/ and /trill-d/ models, only the linear trend was significant, such that a higher vocabulary rating predicted higher forced choice accuracy for the /f-p/ contrast.

Table 7: Summary of logistic regression analysis for /f-p/ contrast

Predictor	B	B		Std Error B	z-value	p
		95% CI				
(Intercept)	2.90	[2.68, 3.14]		0.115	25.186	<.001
Vocab	18.69	[12.60, 24.52]		3.025	6.177	<.001
Vocab^2	-6.43	[-14.09, -0.01]		3.525	-1.824	0.068
Vocab^3	-0.59	[-9.89, 6.55]		4.005	-0.146	0.884
Vocab^4	-6.09	[-15.52, 1.38]		4.143	-1.469	0.142
Vocab^5	1.44	[-5.30, 7.96]		3.325	0.432	0.666

Note.  $\chi^2(5) = 63.14, p < .001$ . B = unstandardized regression coefficient. Numbers in brackets indicate the lower and upper limits of a 95% confidence interval. \*  $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$

### 3.2 Analysis of forced choice accuracy for only well-known words

In order to further investigate the effect of proficiency, the data were divided by group and restricted to only the highest possible vocabulary rating. For a trial to be included for the L2 learners, that participant had to have chosen “I know this word well, can provide a translation in English, and can use this word while speaking Spanish” for the word tested in that trial. These data are represented in Table 8 and Figure 2.

Table 8: Accuracy in the forced choice task for words with highest vocabulary rating only

/tap-trill/		/tap-d/		/trill-d/		/f-p/	
Mean	SD	Mean	SD	Mean	SD	Mean	SD

Intermediate	74.1	15.4	81.5	10.9	90.9	8.7	97.6	3.5
	(48.3-94.3)		(47.3-97.3)		(62.5-100)		(87.9-100)	
Advanced	94.3	11.5	96.3	5.6	98.1	3.2	99.4	1.1
	(59.0-100)		(82.1-100)		(90.0-100)		(97.5-100)	
Spanish NS	97.9	2.4	98.4	2.4	98.7	1.4	99.1	1.3
	(95.1-100)		(94.7-100)		(97.3-100)		(97.5-100)	

Note. All numbers are percentages. Numbers in parentheses show range of scores for each contrast.

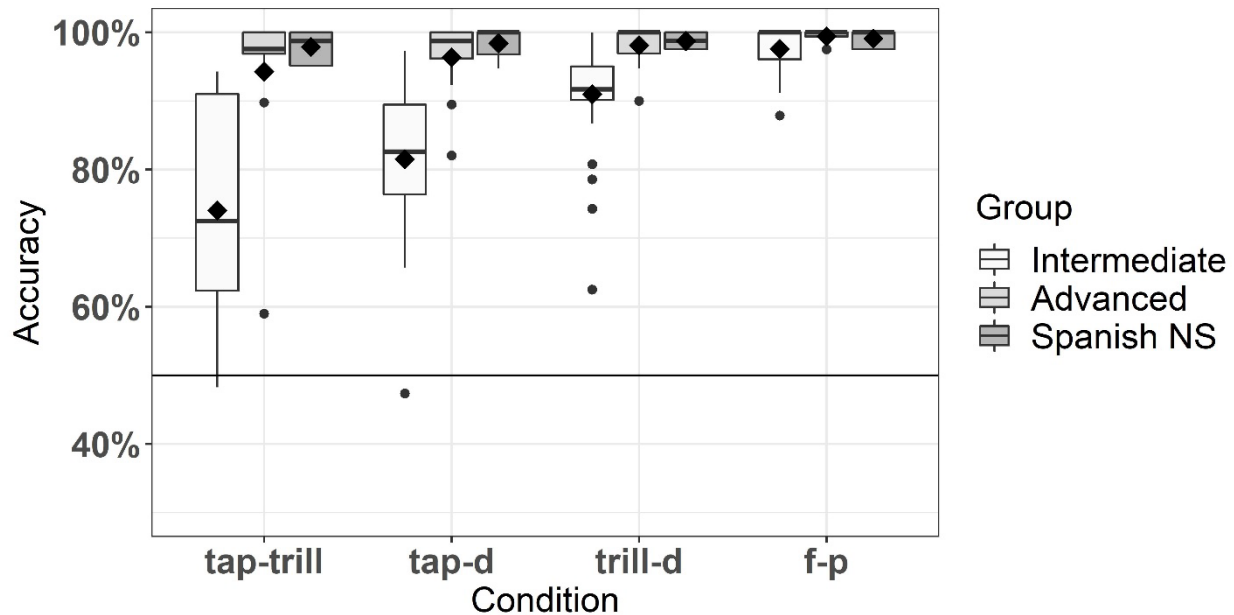


Figure 2: Accuracy in the forced choice task for words with highest vocabulary rating only

A two-way mixed ANOVA was run with *accuracy score* as the dependent variable, *group* (Intermediate, Advanced, or Spanish NS) as the between-subjects independent variable, and *contrast* (/tap-trill/, /tap-d/, /trill-d/, /f-p/) as the within-subjects independent variable. The Bonferroni correction method was used to adjust for multiple comparisons in post-hoc tests. Results showed that there was a significant interaction between group and contrast,  $F(4.36, 93.67) = 8.56, p < .001$ .<sup>2</sup> Post-hoc tests revealed that while there was no effect of group for the /f-p/ control

<sup>2</sup>Because the data violated the assumption of sphericity as shown by Mauchly's Test of Sphericity ( $p < .001$ ), the Greenhouse-Geisser sphericity correction was applied to the degrees of freedom. The data were judged to be approximately normally distributed through an examination of the QQ plot; however, according to the results of Levene's tests, the FCLD data violated the assumption of homogeneity of variance in the /tap-trill/ ( $p < .001$ ), /tap-d/ ( $p = .034$ ), and /trill-d/ ( $p = .044$ ) conditions. Box's M-test revealed that the homogeneity of covariance assumption was additionally violated ( $p < .001$ ), and the data contained two extreme outliers. Therefore, a two-way mixed ANOVA with Robust Estimation was run with the R package WRS2 v.1.0-0 (Mair 2019) following Wilcox (2012). Results mirrored those of the traditional

contrast (adjusted  $p = .52$ ), there were significant differences by group for the other three contrasts (all adjusted  $p < .01$ ). For each of the contrasts /tap-d/, /tap-trill/, and /trill-d/, there was no difference between the advanced learners and the native speakers (all adjusted  $p = 1$ ); however, there were significant differences between the intermediate learners and the other two groups for these contrasts (all adjusted  $p < .05$ ). Contrast was only significant for the intermediate group, for whom all contrasts significantly differed from each other (all adjusted  $p < .05$ ). In sum, while the advanced learners were as accurate as the native speakers across all contrasts, the intermediate learners were significantly less accurate than both the advanced and native speakers for the three test contrasts /tap-d/, /tap-trill/, and /trill-d/. The intermediate learners were least accurate for the /tap-trill/ contrast, followed by the /tap-d/ and /trill-d/ contrasts, respectively. For the /f-p/ control contrast, they were as highly accurate as the advanced learners and the native speakers.

### 3.3 Qualitative analysis of sound-spelling correspondence knowledge

Given that learners' understanding of sound-spelling correspondences could affect how they have phonologically encoded these words, an examination was also carried out on learners' responses to the question in the background questionnaire about the difference in pronunciation between intervocalic <r> and <rr>. Only 2 out of the 12 advanced learners did not give a reasonable explanation of the difference between tap and trill, along with 3 of the 26 intermediate learners. The majority of the advanced learners gave a linguistically-informed explanation of the difference. This was true for a small number of the intermediate learners, although most of them wrote about <rr> being "rolled". Three examples per group that were judged to be accurate are provided below verbatim:

#### Advanced Learners

- (1a) The 'r' as in 'pero' is a tap sound similar to the 'tt' in the North American English word 'better', whereas the 'rr' is a trilled 'r' so the tip of your tongue has to tap your alveolar ridge multiple times
- (1b) 'r' is a tap meaning the tongue hits the alveolar ridge only once, 'rr' is a trill meaning it rapidly hits the alveolar ridge around three times
- (1c) The "rr" involves multiple tongue vibrations while the "r" has only one tongue tap.

#### Intermediate Learners

- (2a) The difference in pronunciation between the two words is there is an emphasis when using a double "rr". One will roll the "rr" to make a trill sound but with the word "pero" you don't trill the "r".

---

two-way mixed ANOVA, with a significant interaction between group and condition and significant main effects of group and condition.

(2b) The single "r" is pronounced with a "tap" of the tongue, whereas the double "rr" is pronounced in a "rolling" manner.

(2c) r is a short sound but rr is more of a rolling sound

The explanations that were judged to be at least partially inaccurate for each group are provided below verbatim:

#### Advanced Learners

(3a) rr is trilled, r is alveolar

(3b) longer trill in the second

#### Intermediate Learners

(4a) The rolled "r" sound

(4b) pero you do roll the 'r' but in perro, you do roll the 'r'

(4c) the trill r is sonora while the r is sordo.

For the advanced learners, the participant who stated that the <rr> represents a “longer trill” (implying that both sounds are trills) was in fact the participant with the lowest accuracy in the forced choice task, at 59%. Nevertheless, the learner who gave the second explanation that <r> is “alveolar” (implying that <rr> is not alveolar, when in reality both sounds have an alveolar place of articulation) did not have a lower score than the rest of the group; their accuracy was 97.5%. For the intermediate learners, the participant who gave the first explanation which does not clarify which spelling corresponds to the “rolled ‘r’ sound” had the 4<sup>th</sup> lowest score among this group, at 56.5%. The participant who gave the second explanation, which may not reflect inaccurate knowledge but instead a typo, scored a 91.4% in the forced choice. Finally, the participant who gave the third explanation, which is wholly inaccurate since both the tap and the trill are voiced, or *sonora*, rather than voiceless, or *sorda*, scored a 75.7% in this task, which was above average.

## 4 Discussion

This study sought to examine whether self-reported knowledge of the meaning of Spanish written words corresponded with accuracy in their phonological form and whether this differed by proficiency level. Although higher vocabulary rating was correlated with higher accuracy in the forced choice task, this association was weak for the intermediate learners and non-existent for the advanced learners. This may be due in part to the lack of substantial variation in vocabulary ratings, which were generally high for most of the words, especially for the advanced learners. Future studies using stimuli for which learners have more varied vocabulary ratings could provide better insight into this relationship.

Despite this limitation, when both groups were combined the regression analyses showed that better knowledge of the meaning of a word does predict a more accurate phonological



representation. This is revealed by the linear relationship for the /tap-trill/, /trill-d/, and /f-p/ contrasts between vocabulary rating and accuracy in the forced choice task. Thus, in general, learners were more accurate at choosing the correct pronunciation of a word the higher they rated their knowledge of that word, although for /f-p/ this trend appears to be driven by differences between unknown words and all other levels of knowledge, rather than a more gradual increase in performance as knowledge increased, as shown by /tap-trill/ and /trill-d/. A linear increase in accuracy was not found for the /tap-d/ contrast, for which the quadratic function, or U-shaped curve, was also significant, signifying that learners' accuracy actually decreased for some of the higher ratings before increasing again. This could be due to several different possibilities. It may be that the gradations in knowledge that the scale was attempting to capture were not meaningful or well understood by the participants, and thus the scale would need to be reworded. It may also be that differences in how well a learner knows the meaning of a word do not appreciably impact the accuracy of its phonological form, and it is the ends of the scale, that is, simply knowing a word exists and being able to use it when speaking, that are the most useful metrics. However, these possibilities would not explain why only the /tap-d/ contrast showed this non-linear trend while the other contrasts did not. A more likely explanation may stem from the perceptual difficulty of this contrast. Studies have shown that the /tap-d/ contrast is the most difficult of the contrasts tested here (Daidone and Darcy 2014, 2021; Rose 2010; Herd 2011). For example, Herd (2011) found that intermediate learners of Spanish had 66% accuracy in identifying tap versus /d/ words after training, compared to 89% for /tap-trill/ and 97% for /trill-d/. Additionally, the nature of [ɾ] as an allophone of /d/ in English makes the Spanish tap and /d/ seem more similar in perception for learners, regardless of their actual acoustic characteristics (Boomershine et al. 2008). Because of the difficulty of /tap-d/, learners may struggle with hearing the difference between these options regardless of their familiarity with the word. Perhaps those learners who were not very familiar with a word containing one of these sounds listened more carefully for the correct form, while learners who were somewhat more secure in their knowledge of a word paid less attention to this difference, and those learners who chose that they can use this word while speaking Spanish were in fact better at recognizing this subtle difference.

When only words that were rated as “I know this word well, can provide a translation in English, and can use this word while speaking Spanish” were included for each participant, accuracy on the forced choice task still ranged greatly by individual for the intermediate group, who scored lower than both the advanced learners and native speakers for all of the test contrasts. They were capable of accurately completing the task when only phonemes that exist in their native language were tested, that is, /f-p/, but struggled to choose the correct pronunciation when tested on new L2 phonemes in the /tap-trill/, /tap-d/, and /trill-d/ contrasts. This is despite the accurate explicit knowledge of the sound-spelling correspondence for intervocalic tap and trill that the vast majority of learners displayed, and for those who did not, they were not uniformly less accurate. Even when knowing how a word should be pronounced based on its spelling and indicating that they could use this word when speaking Spanish, most intermediate learners could not select the correct pronunciation when given a clear choice between two options. Thus, knowledge of the meaning and spelling of a word do not appear to be sufficient for an accurate phonological representation. Moreover, accuracy appears to be driven more by proficiency level than the knowledge of individual words, as the advanced speakers were as accurate as the native speakers.

These findings corroborate previous research that reported L2 phonological representations to be inaccurate or less detailed (e.g. Barrios and Hayes-Harb 2021; Cutler, Weber and Otake 2006; Daidone and Darcy 2014; Gor et al. 2021), and additionally this study shows that the difficulties learners have with storing accurate phonological representations cannot be explained away by the type of task used or participants' insufficient self-reported knowledge of the items tested.

## 5 Conclusions

These results have implications for both vocabulary teaching and testing. In the classroom, instructors cannot assume that knowing how a word is supposed to sound based on its spelling will lead to accurate representations of the sounds within that word. Even Spanish majors and minors, the intermediate learners in this study, need more training on perception and pronunciation when learning vocabulary, and they also could benefit from instruction on what is not an acceptable pronunciation of a word. This is true even for a language with a transparent orthography like Spanish. Moreover, if learners are only assessed on the written form of new vocabulary words, these tests may be missing critical information about the phonological accuracy of learners' vocabulary knowledge. The results of the intermediate learners for words well-known in writing indicate that if researchers are interested in learners' aural vocabulary abilities, the development of aural vocabulary tests for L2 learners or adaptation of such existing tests already in place for children is warranted, even for a language like Spanish. Furthermore, researchers should use caution if extrapolating phonological properties of L2 learners' lexicons such as phonological neighborhood density based solely on written tests. For example, if learners do not have a clear distinction between the Spanish tap and trill, their phonological forms for words such as *pero* 'but' and *perro* 'dog' may be merged, meaning that these words are not phonological neighbors of each other in their mental lexicons but rather homophones, or these representations may be separate but fuzzy to the extent that they are activated by similar words and each other.

Overall, these results reveal a disconnect between self-rated vocabulary knowledge and the accuracy of phonological forms. While having greater knowledge of a word predicts that it is more likely to have an accurate phonological representation, an accurate form is not guaranteed even when a word is well known. This is particularly true for intermediate learners, who often struggled to identify the real word when given a binary choice. Despite the relatively transparent orthography of Spanish, it is clear that phonological forms cannot be taken for granted based on knowledge of written forms. Additionally, it does not appear that learners' difficulty can be attributed to their lack of knowledge of sound-spelling correspondences. Thus, more time needs to be devoted to training the phonological forms of words in the classroom, and both instructors and researchers should use caution when using solely written instruments to assess learners' vocabulary knowledge, as they are likely to overestimate the accuracy of learners' phonological knowledge of these words.

## References

Barrios, Shannon & Rachel Hayes-Harb. 2021. L2 processing of words containing English /æ/-

- /ɛ/ and /l-/ɜ/ contrasts, and the uses and limits of the auditory lexical decision task for understanding the locus of difficulty. *Frontiers in Communication* 6(689470) doi:10.3389/fcomm.2021.689470.
- Beglar, David & Paul Nation. 2014. Assessing vocabulary. In Antony John Kunnan (ed.), *The companion to language assessment*, 172–184. Hoboken, NJ: John Wiley & Sons. doi:10.1002/9781118411360.wbcla053.
- Boomershine, Amanda, Kathleen Currie Hall, Elizabeth Hume & Keith Johnson. 2008. The influence of allophony vs. contrast on perception: The case of Spanish and English. In Peter Avery, B Elan Dresher & Keren Rice (eds.), *Contrast in phonology: Perception and acquisition*, 145–171. Berlin, Germany: Mouton.
- Carney, Nathaniel. 2021. Diagnosing L2 listeners' difficulty comprehending known lexis. *TESOL Quarterly* 55(2). 536–567. doi:10.1002/tesq.3000.
- Cook, Svetlana V., Nick B. Pandža, Alia K. Lancaster & Kira Gor. 2016. Fuzzy nonnative phonolexical representations lead to fuzzy form-to-meaning mappings. *Frontiers in Psychology* 7. 1–17. doi:10.3389/fpsyg.2016.01345.
- Cutler, Anne, Andrea Weber & Takashi Otake. 2006. Asymmetric mapping from phonetic to lexical representations in second-language listening. *Journal of Phonetics* 34(2). 269–284.
- Daidone, Danielle & Isabelle Darcy. 2014. Quierro comprar una guitarra: Lexical encoding of the tap and trill by L2 learners of Spanish. In Ryan T. Miller, Katherine I. Martin, Chelsea M. Eddington, Ashlie Henery, Nausica Marcos Miguel, Alison M. Tseng, Alba Tuninetti & Daniel Walter (eds.), *Selected Proceedings of the 2012 Second Language Research Forum: Building Bridges between Disciplines*, 39–50. Somerville, MA: Cascadilla Proceedings Project.
- Daidone, Danielle & Isabelle Darcy. 2021. Vocabulary size is a key factor in predicting second language lexical encoding accuracy. *Frontiers in Psychology* 12(688356) doi:10.3389/fpsyg.2021.688356.
- Darcy, Isabelle, Danielle Daidone & Chisato Kojima. 2013. Asymmetric lexical access and fuzzy lexical representations in second language learners. *The Mental Lexicon* 8(3). 372–420. doi:10.1075/ml.8.3.06dar.
- Darcy, Isabelle & Trisha Thomas. 2019. When blue is a disyllabic word: Perceptual epenthesis in the mental lexicon of second language learners. *Bilingualism: Language and Cognition* 22(5). 1141–1159. doi:https://doi.org/10.1017/S1366728918001050.
- Dunn, Lloyd M. & Leota M. Dunn. 1997. *Peabody Picture Vocabulary Test - Third Edition*. Circle Pines, MN: American Guidance Service.
- Durham, Kristie, Rachel Hayes-Harb, Shannon Barrios & Catherine E. Showalter. 2016. The influence of various visual input types on L2 learners' memory for phonological forms of newly-learned words. In John M. Levis, Huong Le, Ivana Lucic, Evan Simpson & Sonca Vo (eds.), *Proceedings of the 7th Pronunciation in Second Language Learning and Teaching Conference*, 98–107. Ames, IA.
- Escudero, Paola, Rachel Hayes-Harb & Holger Mitterer. 2008. Novel second-language words

- and asymmetric lexical access. *Journal of Phonetics* 36(2). 345–360.
- Escudero, Paola, Ellen Simon & Karen E. Mulak. 2014. Learning words in a new language: Orthography doesn't always help. *Bilingualism* 17(2). 384–395. doi:10.1017/S1366728913000436.
- Gor, Kira, Svetlana Cook, Denisa Bordag, Anna Chrabaszcz & Andreas Opitz. 2021. Fuzzy lexical representations in adult second language speakers. *Frontiers in Psychology* 12(732030) doi:10.3389/fpsyg.2021.732030.
- Hayes-Harb, Rachel, Kelsey Brown & Bruce L. Smith. 2018. Orthographic input and the acquisition of German final devoicing by native speakers of English. *Language and Speech* 61(4). 547–564. doi:10.1177/0023830917710048.
- Hayes-Harb, Rachel & Hui Wen Cheng. 2016. The influence of the Pinyin and Zhuyin writing systems on the acquisition of Mandarin word forms by native English speakers. *Frontiers in Psychology* 7(785) doi:10.3389/fpsyg.2016.00785.
- Hayes-Harb, Rachel & Jane Hacking. 2015. The influence of written stress marks on native English speakers' acquisition of Russian lexical stress contrasts. *The Slavic and East European Journal* 59(1). 91–109.
- Hayes-Harb, Rachel, Janet Nicol & Jason Barker. 2010. Learning the phonological forms of new words: Effects of orthographic and auditory input. *Language and Speech* 53(3). 367–381.
- Herd, Wendy. 2011. *The perceptual and production training of /d,r,r/ in L2 Spanish: Behavioral, psycholinguistic, and neurolinguistic evidence*. Lawrence, KS: University of Kansas dissertation.
- Herd, Wendy, Joan Sereno & Allard Jongman. 2015. Cross-modal priming differences between native and nonnative Spanish speakers. *Studies in Hispanic and Lusophone Linguistics* 8(1). 135–155. doi:10.1515/shll-2015-0005.
- Hualde, José Ignacio. 2005. *The sounds of Spanish*. New York, NY: Cambridge University Press.
- Izura, Cristina, Fernando Cuetos & Marc Brysbaert. 2014. Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size. *Psicológica* 35(1). 49–66.
- Kaplan, Edith F., Harold Goodglass & Sandra Weintraub. 2001. *The Boston Naming Test*. 2nd edn. Philadelphia, PA: Lippincott Williams & Wilkins.
- Khamis, Harry. 2008. Measures of association: How to choose? *Journal of Diagnostic Medical Sonography* 24(3). 155–162. doi:10.1177/8756479308317006.
- Ladefoged, Peter & Keith Johnston. 2011. *A Course in Phonetics*. 6th edn. Boston, MA: Wadsworth.
- Leeuw, Joshua R. de. 2015. jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods* 47(1). 1–12. doi:10.3758/s13428-014-0458-y.
- Llompарт, Miquel. 2021. Phonetic categorization ability and vocabulary size contribute to the encoding of difficult second-language phonological contrasts into the lexicon. *Bilingualism:*

- Language and Cognition* 24(3). 481–496. doi:<https://doi.org/10.1017/S1366728920000656>.
- Mair, Patrick. 2019. WRS2: A collection of robust statistical methods. R package version 1.0-0.
- Masrai, Ahmed. 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. doi:10.1515/applirev-2018-0106.
- Mathieu, Lionel. 2016. The influence of foreign scripts on the acquisition of a second language phonological contrast. *Second Language Research* 32(2). 145–170. doi:10.1177/0267658315601882.
- McLean, Stuart, Brandon Kramer & David Beglar. 2015. The creation and validation of a listening vocabulary levels test. *Language Teaching Research* 19(6). 741–760. doi:10.1177/1362168814567889.
- Meara, Paul M. 2005. X\_Lex: The Swansea Vocabulary Levels Test. Swansea, UK: Lognostics.
- Melnik, Gerda Ana & Sharon Peperkamp. 2019. Perceptual deletion and asymmetric lexical access in second language learners. *The Journal of the Acoustical Society of America* 145(1). EL13–EL18. doi:10.1121/1.5085648.
- Milton, James. 2009. *Measuring second language vocabulary acquisition*. Bristol, UK: Multilingual Matters.
- Milton, James & Nicola Hopkins. 2006. Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners. *Canadian Modern Language Review* 63(1). 127–147. doi:10.3138/cmlr.63.1.127.
- Nation, Paul & David Beglar. 2007. A vocabulary size test. *The Language Teacher* 31(7). 9–13.
- Rafat, Yasaman. 2016. Orthography-induced transfer in the production of English-speaking learners of Spanish. *The Language Learning Journal* 44(2). 197–213. doi:10.1080/09571736.2013.784346.
- Ramus, Franck, Sharon Peperkamp, Anne Christophe, Charlotte Jacquemot, Sid Kouider & Emmanuel Dupoux. 2010. A psycholinguistic perspective on the acquisition of phonology. In Cécile Fougerson, Barbara Kuehnert, Mariapaola Imperio & Nathalie Vallee (eds.), *Laboratory Phonology 10: Variation, Phonetic Detail and Phonological Representation*, 311–340. Berlin, Germany: De Gruyter Mouton. doi:<https://doi.org/10.1515/9783110224917>.
- Reilly, Jamie, Vanessa Troiani, Murray Grossman & Arthur Wingfield. 2007. An introduction to hearing loss and screening procedures for behavioral research. *Behavior Research Methods* 39(3). 667–672.
- Rose, Marda. 2010. Differences in discriminating L2 consonants: A comparison of Spanish taps and trills. In Matthew T. Prior, Yukiko Watanabe & Sang-Ki Lee (eds.), *Selected proceedings of the 2008 Second Language Research Forum*, 181–196. Somerville, MA: Cascadilla Proceedings Project.

- Rose, Marda. 2012. Cross-language identification of Spanish consonants in English. *Foreign Language Annals* 45(3). 415–429. doi:10.1111/j.1944-9720.2012.01197.x.
- Schmitt, Norbert, Diane Schmitt & Caroline Clapham. 2001. Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing* 18(1). 55–88. doi:10.1177/026553220101800103.
- Showalter, Catherine E. & Rachel Hayes-Harb. 2013. Unfamiliar orthographic information and second language word learning: A novel lexicon study. *Second Language Research* 29(2). 185–200. doi:10.1177/0267658313480154.
- Showalter, Catherine E. & Rachel Hayes-Harb. 2015. Native English speakers learning Arabic: The influence of novel orthographic information on second language phonological acquisition. *Applied Psycholinguistics* 36(1). 23–42. doi:10.1017/S0142716414000411.
- Simon, Ellen, Della Chambless & Ubiratã Kickhöfel Alves. 2010. Understanding the role of orthography in the acquisition of a non-native vowel contrast. *Language Sciences* 32(3). 380–394. doi:10.1016/j.langsci.2009.07.001.
- Vitevitch, Michael S., Melissa K. Stamer & Douglas Kieweg. 2012. Short research note: The beginning Spanish lexicon: A web-based interface to calculate phonological similarity among Spanish words in adults learning Spanish as a foreign language. *Second Language Research* 28(1). 103–112. doi:https://doi.org/10.1177/0267658311432199.
- Wesche, Marjorie & T. Sima Paribakht. 1996. Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review* 53(1) doi:10.3138/cmlr.53.1.13.
- Wilcox, Rand. 2012. *Modern statistics for the social and behavioral sciences: A practical introduction*. New York, NY: Taylor & Francis.