

Influence of orthography in production and perception of /b/ in US Spanish

Annie Helms^a, Gabriella Licata^b, Rachel Weiher^c

^a University of California, Berkeley (USA), annie_helms@berkeley.edu

^b University of California, Berkeley (USA), glicata@berkeley.edu

^c University of California, Berkeley (USA), rweiher@berkeley.edu

Article history

Received: 13/08/2021

Accepted: 20/01/2022

Keywords

Perception

Bilingualism

Allophones

Orthography

US Spanish

This paper examines the effect of orthography and language profile on Spanish-English bilinguals' production and perception of intervocalic /b/. We hypothesize more labiodental productions of /b/ and weakened discrimination of allophones to be correlated with the grapheme <v>. We also hypothesize that early bilinguals will have more labiodental productions and weaker discrimination of the allophones. Results of a production and discrimination task indicate that <v> is correlated to higher relative intensity in the production task and lower discrimination accuracy in the perception task, regardless of the degree of exposure to English. These results advocate for a usage-based model of language representation.

1. Introduction

In prescriptive Spanish phonology, labiodental productions of /b/ (i.e., more [v]-like) have been historically unattested. However, recent studies have demonstrated evidence for labiodental productions as a relic variant (Figueroa Candia & Evans, 2021; Sadowsky, 2010; Torres Cacoullos & Ferreira, 2000; Vergara & Pérez, 2013), as a marker of emphatic speech (Lope Blanch, 1988), and as a result of language contact (Chappell, 2019; Torres Cacoullos & Ferreira, 2000; Trovato, 2018). Beyond mere observation of /b/ allophones, several studies have explored various influences on the production and discrimination of variants, including token frequency (Torres Cacoullos & Ferreira, 2000), acoustic metrics (Trovato, 2018), phonetic environment (Sadowsky, 2010; Trovato, 2018), English cognate status (Torres Cacoullos & Ferreira, 2000), and orthography (Torres Cacoullos & Ferreira, 2000; Trovato, 2018), among others. In line with bilingual and usage-based models of language representation, we maintain that Spanish language contact with English—the latter of which includes phonemic /v/ in its phonological inventory—is a principal source of influence to account for the presence of labiodental variants in

US Spanish. We examine the production and perception of intervocalic /b/ in Spanish by L1-Spanish L2-English bilinguals across language experience and orthographic representation of /b/. Based on the demonstrated connection between language experience, production, and perception (Flege, 1995; Flege & Bohn, 2021; Johnson, 1997), we expect production to be correlated with perception, whereby early bilinguals produce more labiodental variants of /b/ and have weakened discrimination of variants. Furthermore, orthographic representations have been shown to influence the perception of phonetic information (Han & Choi, 2016, p. 758; Rafat, 2015). Accordingly, we predict orthographic <v> to be correlated with more labiodental productions and less accurate discrimination among allophones. In employing a two-pronged experiment to assess the production and perceptual discrimination of a continuum of variants between [β] and [v] by early and late bilingual speakers (L1-Spanish L2-English) in the United States, the present experiment contributes to a gap in the literature by exploring the link between production and perception and how language experience and orthography may influence how sounds are organized and represented by the bilingual speaker-listener.

The paper is organized as follows: first, we provide a review of the literature, in which we discuss theories accounting for labiodental productions of /b/ in Spanish and provide an acoustic description of the range of /b/ allophones. We also discuss the role of orthography and language experience in production and perception, as well as the relationship between production and perception. Subsequently, we detail our methodology, including information about participants, creation of stimuli, experimental procedure, analysis, and statistical modeling. Lastly, we present our results along with a discussion of their implications for speech production and perception among Spanish-English bilinguals, as well as for bilingual speakers more generally.

2. Literature review

2.1. Theories to account for [v]

Until the 15th century, Spanish had two distinct voiced labial consonants, /b/ and /v/, represented orthographically by and <v> respectively (Penny, 2002, p. 85). These distinct sounds underwent a process of leveling, whereby /b/ and /v/ neutralized to [β] in certain linguistic contexts (i.e., intervocalically). Traditional phonological accounts describe [b] and [β] as being in complementary distribution in Spanish, with [b] occurring in utterance-initial position and after nasals and [β] occurring elsewhere (p. 86). Due to a rather small number of minimal pairs (e.g., *savia* ‘sap’ and *sabia* ‘wise’) and little chance of miscommunication, this neutralization to [β] became widespread in Spanish. However, the voiced labiodental fricative [v] has been observed in some varieties of modern Spanish (Chappell, 2019; Ortega, 2018; Sadowsky, 2010; Torres Cacoullós & Ferreira, 2000; Trovato, 2018; Vergara & Pérez, 2013). Following Lope Blanch (1988) and Torres Cacoullós and Ferreira (2000), among others, the three accounts for voiced labiodental fricative [v] productions of /b/ in Spanish are: (1) residual presence of medieval Spanish phone /v/, formalized as Archaic Theory; (2) an articulatory consequence of emphatic speech and hypercorrection; and (3) a result of language contact with languages possessing a voiced labiodental fricative in their phonological inventories (e.g., English, Balearic Catalan, Portuguese).

Archaic Theory (Alonso, 1967) stipulates that some varieties of Spanish maintain the phone /v/ as an underlying form, thereby yielding voiced labiodental fricative [v] productions. The evolution of the voiced bilabial stop /b/ from Latin to Modern Spanish included intermediate labiodentalization, which was neutralized in Modern Spanish when the default place of articulation became bilabial (Alonso, 1967; Martínez-Gil, 1998). First, /b/ underwent spirantization after a nuclear vowel, then subsequent labiodentalization. Next, the context in which spirantization and subsequent labiodentalization could occur widened to include a position following a [+continuous] segment. Surface voiced labiodental fricative [v] was reanalyzed as underlying /v/ and voiceless bilabial stop /p/ was voiced and subsequently reanalyzed as underlying /b/. This bilabial phone again underwent spirantization and labiodentalization following a nuclear vowel or [+continuous] segment. Underlying /v/ was strengthened to [b] after pauses and nasals, then subsequently reanalyzed to /b/. In most varieties of Spanish, the default place of articulation in all contexts became bilabial, where [b] is produced after pauses and nasals and [β] is produced in other contexts. Archaic Theory relies on this progression to account for the labiodental productions of /b/ in some varieties of Spanish, where the last step of the default bilabial place of articulation, merging /b/ and /v/, did not occur (Lope Blanch, 1988; Torres Cacoullós & Ferreira, 2000; Trovato, 2018). Instead, speakers of these varieties have two underlying forms /b/ and /v/, resulting in respective surface forms of [b] after a pause and after a nasal and [v] in all other contexts. As the bilabial/labiodental distinction is on the phonological level rather than the orthographic level, speakers of these varieties of Spanish produce labiodentals regardless of orthographic representation (Sadowsky, 2010; Vergara & Pérez, 2013).

An alternative explanation for the presence of labiodental productions of /b/ is that they are a phonetic resource to convey emphatic speech, especially in Mexican Spanish (Mexican pedantic [v]; Lope Blanch, 1988). As such, labiodental productions are most frequent in formal settings, rather than spontaneous speech, and are associated with a pedantic style of speech. Orthography is a strong predictor of labiodental production, as the phenomenon is associated with hypercorrection,

where labiodental productions only occur sporadically with the grapheme (Lope Blanch, 1988, p. 169). Labiodental productions are attested in all phonetic contexts, both where [b] and [β] prescriptively correspond, although a preceding consonant appears to be most favorable.

Lastly, the third explanation in the literature for labiodental /b/ is language contact (Chappell, 2019; Torres Cacoullós & Ferreira, 2000; Trovato, 2018). Under this account, speakers of Spanish that are bilingual in a language with a /v/ phone (e.g., Balearic Catalan, English, Portuguese) may produce more labiodental-like productions of Spanish /b/ due to the presence of /v/ in their bilingual phonetic repertoire. Both the Speech Learning Model (SLM; Flege, 1995) and the Revised Speech Learning Model (SLM-r; Flege & Bohn, 2021) postulate that if phonetic differences between an L2 category and an L1 category are not perceived by a bilingual individual, the formation of a new category will be blocked. Blockage of category formation results in assimilation, by which the speaker may produce a composite L1-L2 category, allowing for acoustic and articulatory properties of an L2 category (such as labiodentalization) to appear in the L1, and vice versa.

2.2. Acoustic description of [β] and [v]

According to prescriptive norms, /b/ spirantizes to [β] when following a [+continuous] segment and [b] only occurs after nasal consonants or after a pause. However, in some varieties of Central American Spanish and in Highland Colombian Spanish, [b] is in free variation with [β], or can even be the preferred variant, after any consonant or semivowel (Canfield, 1981; Carrasco et al., 2012; Lipski, 1994). Notably, even in these varieties, /b/ is spirantized in postvocalic position. In recognition of the existence of variable spirantization across varieties of Spanish, the present study exclusively analyzes /b/ in intervocalic and posttonic position, where /b/ is most likely to be spirantized and not produced as a plosive (Ortega-Llebarria, 2003). In the review of the literature that follows, we assume that possible variants of /b/ fall on the allophonic continuum that includes approximant [β], plosive [b] and fricative [v]. However, we at times draw specific contrasts between [β] and [v], between [β]/[b] and [v], or between [β] and [b], depending on

the component(s) of contrast that we wish to highlight.

Keeping in line with the language contact theory of labiodentalization in Spanish, recent studies in phonetic production have addressed the variation of /b/ in US Spanish through empirically grounded accounts (Rao, 2014; Torres Cacoullós & Ferreira, 2000; Trovato, 2018). Results from these studies yield a range of allophones from [β] to [v], which are gradient when measured by the acoustic properties, e.g., relative intensity, duration, and spectral moments. The acoustic properties of bilabial approximants and labiodental fricatives are relatively similar due to place of articulation; however, they remain fairly understudied (Trovato, 2018, p. 29), partly due to their rarity within a language variety (Ladefoged & Maddieson, 1996) but also in part due to their variability in phonetic realization (see Figure 1). Bilabial fricatives (as well as approximants, cf. Hualde, 2013) involve movement of both the upper and lower lips, while labiodental fricatives in large part only require movement of the lower lip (Ladefoged & Maddieson, 1996, p. 140). Both allophones exhibit varying degrees of frication, though it is more apparent at the midpoint of [v] due to more prolonged contact between the upper teeth and lower lip. Additionally, the duration of [β] is substantially shorter than that of [v], though approximants like [β] can be difficult to measure due to less defined transitions to following vowels due to a lack of articulatory contact. Center of gravity (henceforth COG), to our knowledge, has only been used by Trovato (2018) to distinguish bilabial and labiodental allophones of /b/, where [β]/[b] may have lower COG than [v] due to its more anterior point of articulation (Stevens, 1960). Mazzaro (2011) found that COG was the best predictor of place of articulation for bilabial and velar fricatives in Argentinian Spanish, providing evidence that COG may be useful for the place of articulation distinction, though with the distance between a bilabial and labiodental articulation, the utility of COG is not expected to be great. The difference in articulation via lip movement can be measured with electromagnetic articulography (EMA) and accordingly can be used to categorize allophonic production (Ladefoged & Maddieson, 1996), though parameters have not yet been defined for these allophonic categories in Spanish. Sadowsky (2010) and Trovato (2018) utilized a visual analysis (cf.

Ladefoged, 2003, p. 32) to detect articulatory differences between bilabials and labiodentals, finding that plain sight was more successful in

detecting the differences among /b/ allophones rather than only relying on auditory and acoustic analyses.

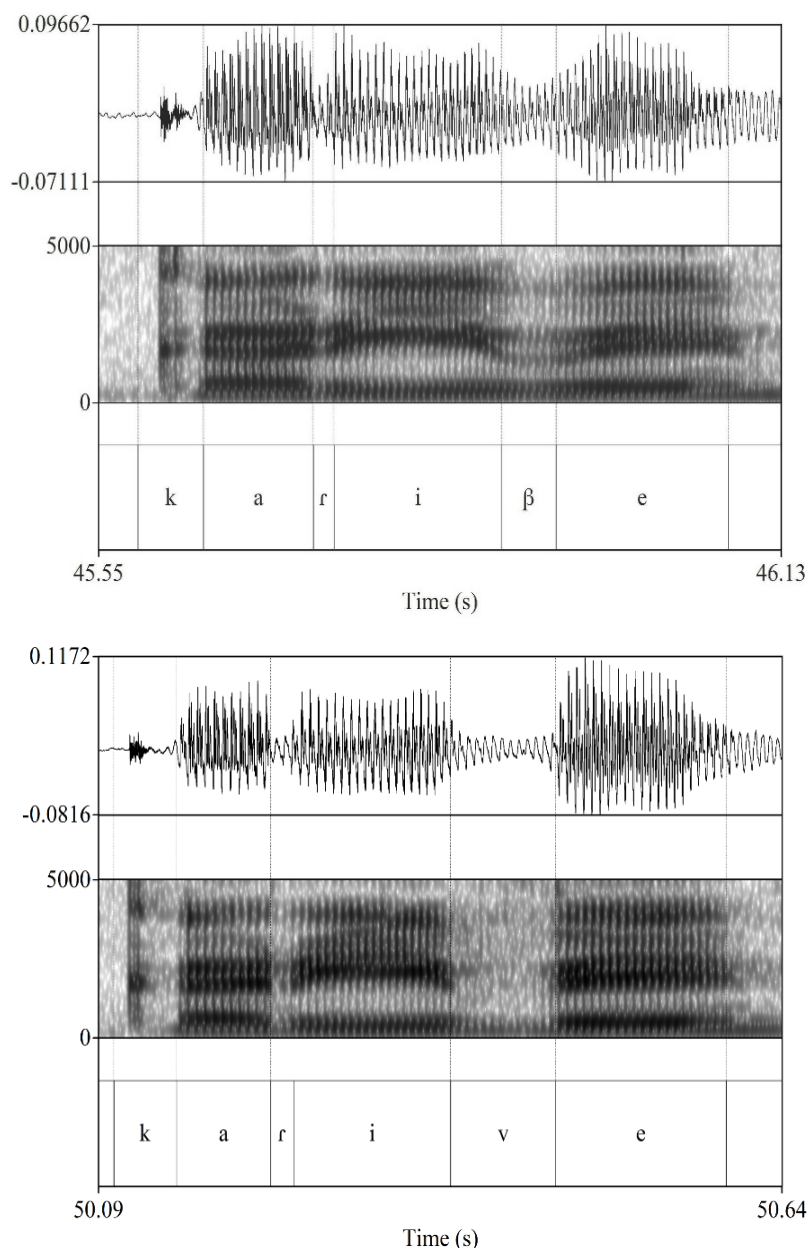


Figure 1. Spectrograms of kari[β]e (above) and kari[v]e (below).

Trovato (2018) coded tokens of /b/ auditorily and visually as having either bilabial or labiodental places of articulation, and examined whether the acoustic measures of duration, COG, and intensity can be used to document the differences between /b/ variation in El Paso Spanish. The differences in articulation between [β]/[b] and [v] lend support to the hypotheses that the latter is produced with longer duration, higher intensity difference, and higher COG (Trovato, 2018, pp. 29-32). In this study,

relative intensity was a significant predictor of perceived place of articulation in all phonetic contexts, where consonants perceived as labiodental were produced with higher relative intensity. Additionally, duration was significant in word-initial position, where perceived labiodental tokens were longer.

Amengual (2019) and Carrasco et al. (2012) used intensity difference (dB; [following vowel

maximum intensity] – [consonant minimum intensity]) to measure the degree of lenition of /b/. Amengual (2019) examined phrase-initial and intervocalic stop production of early sequential, simultaneous, and later-learner bilinguals. He found that the first group lenited intervocalic segments more than the latter two, demonstrating the formidability of phonetic exposure in the early years of language learning. Carrasco et al. (2012) demonstrate that patterns of voiced stop allophony differ significantly between Spanish dialects. They confirmed that in terms of constriction, speakers in Costa Rica clearly distinguish stop production in post postconsonantal position from that of postvocalic, while speakers in Madrid, Spain exploit a continuum of constriction degrees that depends on several linguistic contexts. Colantoni and Marinescu (2010) assessed the correlation between voiced and voiceless stop weakening in intervocalic position in Argentine Spanish. To determine the degree of lenition, they measured relative intensity and duration, which would respectively increase and shorten if stops were weakened. The authors found that voiced and voiceless stop weakening was not correlated, as voiced stop lenition (i.e., spirantization) was blocked in stressed syllables (p. 113). Eddington (2011) measured the relative intensity of Spanish approximants [β ð γ] from a variety of Spanish dialects and in a variety of linguistic contexts and found [β] is most lenited intervocalically, a position that thus decreases its relative intensity.

As noted earlier, there is scant research that assesses listener perceptions of /b/ allophones nor any research that considers a listener's language profile or acoustic measures to potentially affect perception. However, a recent study on the perception of approximants [β] and [v] in Chilean Spanish (Figueroa Candia & Evans, 2021) found that listeners do not categorically discriminate the two sounds, providing evidence that these allophones are not sociolinguistically salient in Chile. However, in two attitudinal perception experiments in both central Mexico and Texas, Chappell (2019, 2020) found that perceptions of the labiodental variant and its orthographic representation were in fact salient, significantly conditioned by guise gender, whereby listeners perceived [v] as a positive social index for female speakers, but as a negative index for male speakers. In this study, we aim to contribute to this growing

body of literature by incorporating orthography and listener language profile to better understand the production and perception of /b/ allophones in US Spanish.

2.3. SLM and SLM-r

Within models of bilingual phonetic representation, such as the SLM and SLM-r, the relationship between perception and production is not clear-cut. Although the SLM and SLM-r both hypothesize that the perception of acoustic differences between two sounds is correlated to an individual's ability to produce the two sounds with acoustic distinction, the SLM posits that the relationship is unidirectional, where production evolves to match perception. In the SLM-r, however, Flege and Bohn (2021) acknowledge a bidirectional relationship between perception and production, a theory that is more in line with usage-based models of language (Bybee, 2000; Johnson, 1997; Pierrehumbert, 2001). Additionally, the authors mention that “[t]he presence of near-mergers, that is, the systematic production of differences that cannot be readily perceived (e.g., Labov, 1994), indicates that production and perception are not completely symmetrical.” (p. 29). Therefore, studies that incorporate both elements of production and perception are crucial to the understanding of bilingual phonetic representation and the cognitive factors that facilitate such processes.

2.3.1. Language Exposure

The acoustic profile of a sound in the L2 which becomes perceptually linked to a sound in the L1 is influenced by several factors, such as the quality and quantity of language input and the relative language activation at the time of production or perception. Relative language input is often conceptualized as language exposure or language dominance and is shaped by factors such as linguistic history, linguistic attitudes, language proficiency, and language use (Gertken et al., 2014). For example, Ortega (2018) analyzes /b/ production in El Paso, Texas, by intermediate heritage speakers of Spanish, advanced heritage speakers, later bilinguals (acquired English after age of 13), and recent arrivals to the US. The intermediate heritage speakers have the highest rate of labiodental productions, followed by the advanced heritage speakers. These results are in line with predictions

based on language exposure, where these two participant groups are expected to have the most exposure to and proficiency in English. Trovato (2018) finds that greater English proficiency is correlated to higher rates of labiodental production. Similarly, Amengual (2019) found that sequential bilinguals produced more approximant-like intervocalic /b d g/ realizations than simultaneous bilinguals, demonstrating that even a five-year delay in the exposure to and acquisition of English significantly affects phonetic production.

A speaker's age of exposure to phonological information has been shown to later condition performance in discrimination tasks. Sebastián-Gallés et al. (2005) showed that Catalan-Spanish sequential bilinguals performed better on a lexical task in Catalan than Spanish-Catalan sequentials, who began acquiring their L2 around age four. Several studies have also shown that children who receive language exposure at a very young age, despite growing up in a completely different linguistic environment, retain discriminatory abilities in their [passive] L1 (Hyltenstam et al., 2009; Pierce et al., 2014; among others). However, few studies have shown the effects of language exposure on the discrimination of a contrast that is allophonic in one language and phonemic in the other. Accordingly, we seek to understand the effects of language exposure on the production and perceptual discrimination of [v] and [β] in Spanish.

2.3.2 Orthography

In addition to the effects of language exposure, orthography has been investigated as a potential factor influencing speech production and perception. While the body of work that focuses on production is less robust (Han & Choi, 2016, p. 758), likely due to traditional understandings that written language holds less weight on cognitive representation of language than spoken language (Rastle et al., 2011, p. 3), some studies have observed mutual influences between phonological and orthographic systems, notably in the acquisition of novel words in the L1 (Rastle et al., 2011, p. 13) and of L2 phonology (Bassetti et al., 2015; Rafat, 2015). Yet, as a whole, results of these studies often reveal that orthography effects are constrained by factors such as task type (Rastle et al., 2011, p. 12; Roelofs, 2007, p. 36; Zhang & Damian, 2012, p. 276) and word type (i.e., real vs. nonce words;

Temkin Martínez & Müllner, 2016, pp. 5-6). For instance, with respect to task-related constraints, Zhang and Damian (2012, p. 277) found that when stimuli were presented visually, they observed an “inhibitory effect” which was not present when stimuli were presented auditorily. The authors also warn that a potential confounding effect of orthographic and phonological correspondence in Chinese could have arisen, and that orthography could guide production more in languages such as English (Zhang and Damian, 2012, p. 277). In contrast, Temkin Martínez and Müllner (2016) provide evidence for orthography effects on the production of variable spirantization in Modern Hebrew, such that when confronted with nonce words, speakers drew on their knowledge of the orthographic representation and variable spirantization to guide production (p. 6). In other words, while instances of apparent orthographic effects have been observed in the literature, there remains little evidence “that orthographic representations are mandatorily activated in word production” (Zhang & Damian, 2012, p. 273). Moreover, orthography has been found to both aid and impede acquiring target-like phonology (Bassetti et al., 2015, p. 1).

In the case of perception, there have been investigations of orthographic influence on latency in bilingual lexical access (Schwartz et al., 2007) and the processing and storage of allophonic variation in nonce words (Han & Choi, 2016), for example. Examining the nuances of lexical activation among Spanish-English bilinguals, Schwartz et al. (2007, pp. 116, 120) found that effects of orthographic similarity among cognates were deeply intertwined with those of phonological similarity, such that their processing of cognates was slowed only when orthographic similarity corresponded to greater phonological distinctiveness. In a study of native English speakers' acquisition of nonce words, Escudero and Wanrooij (2010) found that speakers' memory of such words was influenced by whether or not the orthography they were shown aligned with typical English grapheme-phoneme relationships (p. 378). Accordingly, this attention to orthographic norms in English evidences a connection, or “contamination” between orthography and perception in language acquisition, even when that language is the L1 (p. 378). Similarly, Han and Choi (2016) examined the storage of nonce lexical items as a locus of the

“restructuring of the preexisting phonological representation,” where they found orthography to ultimately affect subsequent production (p. 771). Additionally, in Modern Hebrew, speakers drew on orthographic representation of spirantization in their production of nonce words, yet their production in real words patterned similarly to their perception (Temkin Martínez & Müllner, 2016, p. 5), thus suggesting a link between perception and production with respect to allophonic variation.

2.4. Research Questions

Studies in bilingual environments that examine the relationship between production and perception, particularly in the context of allophones, are lacking in the prior literature (see Mazzaro & González de Anda, 2019). To this end, we investigate the link between gradient production of Spanish /b/, ranging from bilabial to labiodental, and its perception on a continuum from [β] to [v]¹. The present experiment examines early and late L1-Spanish L2-English bilingual speakers’ production and perception of intervocalic /b/ of Spanish in contact with American English in order to determine: 1) whether orthography is a significant predictor of any of the acoustic metrics correlated with gradient productions; 2) if production differs significantly between early and late bilinguals; 3) how the discrimination of [β] and [v] of speakers with more labiodental productions compares to that of speakers with less labiodental productions; 4) whether discrimination differs based on the orthographic representation; and 5) whether discrimination is significantly different between early and late bilinguals. This experimental methodology contributes to a growing body of empirical research on the ways in which bilingual speakers establish links between the phonetic and orthographic systems of different languages, and how these links bear out in both production and perception.

Based on findings from previous studies (Trovato, 2018), we predict orthography to significantly condition production, where the grapheme <v> is correlated with more labiodental productions of /b/. Following Trovato (2018), we will use relative

duration, COG, and relative intensity as possible correlates of the degree of labiodentalization, where more labiodental productions might have longer duration, higher COG, and higher relative intensity. Due to earlier and more extensive contact with English, we hypothesize that early Spanish-English bilinguals (both simultaneous and sequential) will produce more labiodentalized allophones of /b/ than late bilinguals. We additionally hypothesize a correlation between production and perception, where speakers that produce more labiodentalized allophones of /b/, namely the early bilinguals, will exhibit less accuracy in discrimination than those with more bilabial productions of intervocalic /b/. Lastly, we hypothesize that orthography will be a significant predictor of discrimination, where the accuracy of perceptual discrimination decreases when <v> is present in the token word.

3. Methodology

3.1. Participants

Twenty-two participants are included in the present experiment and consist of two groups: early Spanish-English bilinguals and late Spanish-English bilinguals, with Spanish as the L1 for both groups. Early bilinguals are defined as speakers that acquired English simultaneously and sequentially before the age 5 (in line with Amengual, 2019), and late bilinguals acquired English later in school as a second language (after the age of 18)². All early bilinguals grew up in California. Participants in both groups consist primarily of graduate and undergraduate students, faculty, and staff at the home institution of the researchers, and were recruited through word of mouth and through emails to departmental mailing lists. All participants reported never having any history of speech or hearing disorders. Participant demographics are found in Table 1.

3.2. Stimuli

A list of token words (Table 2) was generated, stratified by orthography (intervocalic and <v>), where intervocalic /b/ occurred in post-tonic

¹ The speaker who produced stimuli for this study uses [β] intervocalically in his linguistic repertoire, rather than the occlusive [b]. He was asked to produce the same tokens using [β] and [v] for the purposes of this experiment. The continua generation process is further detailed in Section 3.2.

² Previous studies (e.g. Amengual, 2019) have demonstrated that simultaneous and sequential bilinguals may pattern differently in phonetic production. Accordingly, future studies could further distinguish early bilinguals into these groups.

positions (Ortega-Llebaria, 2003) and in controlled vocalic contexts. Cognate words were distributed across the factor of orthography and all words had a relative frequency of at least 4 ppm (parts per million) according to the online corpus NIM (Guasch et al., 2013). Cole et al. (1999) have demonstrated that spirantization of intervocalic consonants in Spanish is facilitated by minimal movement of the tongue between production of the consonant and the following vowel. Accordingly, to facilitate the production of [β] (rather than [b]), we limited the surrounding vowel context to /a/, /e/, and

/i/. The speaker who assisted in the generation of the experimental stimuli is a bilingual graduate student in his mid-twenties whose L1 is Spanish and was learned in Mexico, where he was born. He began to learn English at age 11 when he arrived in the United States, where he has lived since. The speaker was recorded reading six repetitions of each token word, three with intervocalic [β] and three with intervocalic [v]. The best repetition of each variant was selected as the basis for the endpoint stimuli

Grew up in:	Language Profile		Gender	
	Early	Late	Female	Male
United States	10	0	7	3
Heritage: Mexico & Central America	8		7	1
Heritage: Caribbean	0		0	0
Heritage: South America	2		0	2
Heritage: Spain	0		0	0
Caribbean	0	1	1	0
Mexico & Central America	0	4	3	1
South America	0	5	4	1
Spain	0	2	2	0
Total:	10	12	17	5

Table 1. Participant demographics ($n = 22$).

Orthographic 	Orthographic <v>
cabe	cadáver
caníbal	comitiva
caribe	detective
casaba	clave
graba	lavan

Table 2. Token stimuli stratified by orthographic representation and in controlled stress and vocalic environments.

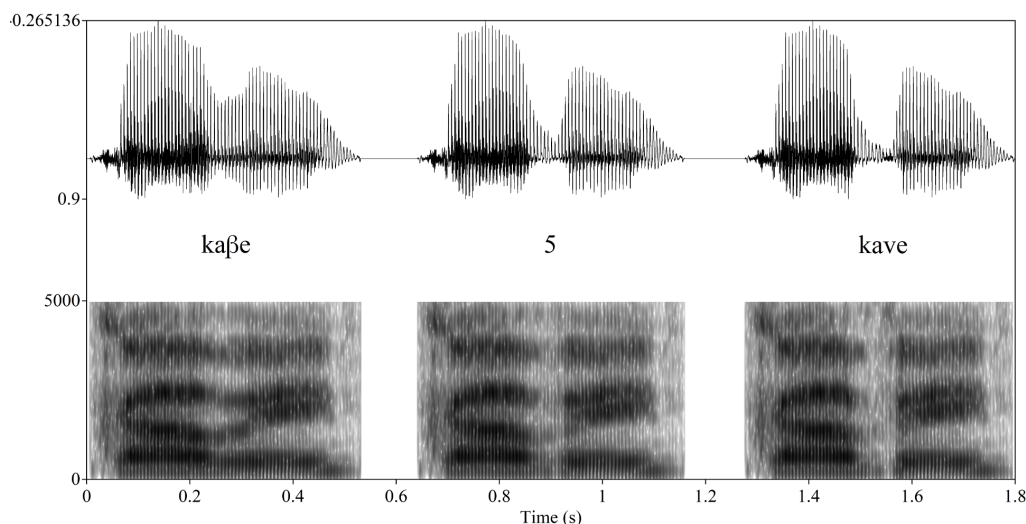


Figure 2. Endpoints and midpoint of 9-step continuum from [β] to [v].

TANDEM-STRAIGHT (Kawahara et al., 2008) is a graphical user interface that generates continua non-parametrically, meaning that steps along the continuum do not differ across only one acoustic metric (e.g., F1). Rather, the spectra of the endpoint tokens are manipulated in their entirety. Using this software, nine-point continua (see Figure 2) were created between the two endpoint stimuli for each word, from which we created trial pairs with a three-step interval. Presently, there have not been studies that concretely categorize each Spanish allophone according to acoustic measurements, thus we were unable to select a step interval based on a standardized acoustic difference. However, a step interval of 3 (i.e., stimuli compared steps of 1-4, 2-5, 3-6, 4-7, 5-8, and 6-9) was chosen to obtain a more precise location of the allophonic boundary to reduce the acoustic difference within the pair and to avoid a ceiling effect from overly obvious discrimination (Gerrits & Schouten, 2004, p. 369). Once the endpoints were established, a base token was generated in *Praat* (Boersma & Weenink, 2019) using the midpoint between the two endpoints, and iterations of /b/ from each step of the continuum were spliced into this base, such that stimuli only varied in the intervocalic segment under investigation, i.e. la[β]an and la[v]an. All splicing occurred at zero crossings to obscure auditory signs of manipulation, and these tokens were played for a native Spanish speaker and rated as sounding natural.

3.3. Experiment design and procedure

3.3.1. Production Task

Prior to the perception task, participants were recorded reading from a randomized list of the 10 token words and 16 filler words, allowing for a comparison of production and perception of /b/ allophones with the orthographic representations

 and <v>. The word list was presented to participants via Microsoft PowerPoint, with a single word on each slide, such that participants were only able to see one word at a time. A total of 220 tokens (10 words x 22 participants) of intervocalic /b/ were obtained using a Zoom H4N Multitrack Recorder.

3.3.2. Perception Task

Pairs of experimental stimuli were concatenated using a *Praat* script (Mayer, 2013) with a 500 ms intrapair interval. Two sets of concatenated pairs, according to the predetermined trial sequence combinations, were then further concatenated with a 1 s interpair interval so that all orders of audio files within the pairs were generated (Figure 3). These interstimulus intervals of 500 ms and 1 s were selected to encourage participants to pay attention to acoustic detail and to lessen the memory load of the task (Gerrits & Schouten, 2004, p. 371). The sets of two concatenated pairs were then organized into a roving 4I2AFC discrimination task (Gerrits & Schouten, 2004), in which only the second and third token of each trial differs from the rest, resulting in sequences of ABAA, AABA, BABB, BBAB, where ‘A’ is always the lower step on the continuum and ‘B’ the higher step. As the endpoints of the continuum are allophones in Spanish, rather than phones, the 4I2AFC discrimination task was selected because this type of task has been demonstrated to elicit continuous perception, rather than categorical perception (Kataoka & Johnson, 2007). Additionally, a 4I2AFC task minimizes uncertainty because participants are only asked to determine *where* a difference exists, not *whether* the difference exists. This resulted in 240 trials, which were counterbalanced and arranged into blocks (to minimize risk of position bias, cf. Wickens, 2001), resulting in two experiments of 120 trials administered via *Qualtrics* (Qualtrics Labs, 2019).

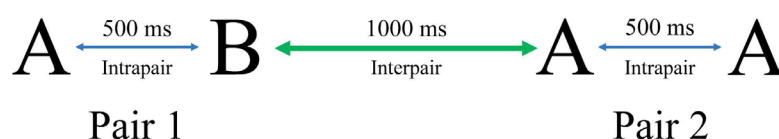


Figure 3. Configuration of experimental tokens

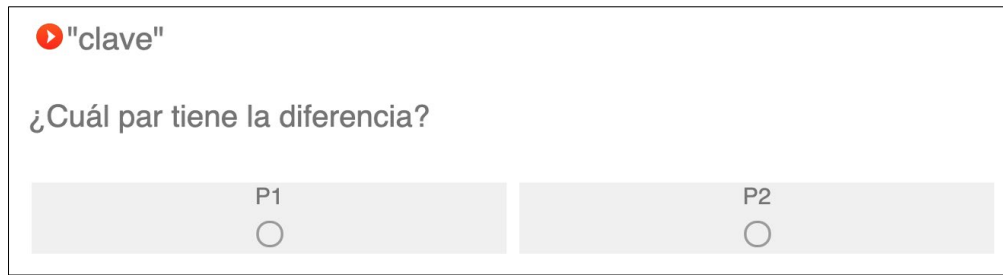


Figure 4. Example of discrimination task in *Qualtrics*.

Participants were given access to the perception task after the production task via a direct link to the questionnaire hosted on Qualtrics. First, they were instructed to complete a word identification task, wherein they typed out the Spanish word they heard in an audio clip, which served to test their hearing and the playback quality of their device. Next, they were presented with a consent form (available in both Spanish and English). If participants failed the word identification task, or if they did not sign the consent form, their session was ended, and they could not complete the experiment. These instructions were written in both English and Spanish to accommodate speaker preference.

Participants were then presented with a practice trial that provided exposure to the experiment design using a different variable ([s] vs [z]). Sound clips were embedded into the questionnaire via Soundcloud and participants were additionally presented with the written form of the token word. The sound clip was set to play automatically, and participants were unable to proceed to the discrimination task question until six seconds had passed, ensuring that they listened to the full audio clip (see Figure 4). They were then able to proceed to the question, “Which pair contains the difference?”. The question appeared in Spanish with two possible answers, P1 (Pair 1) or P2 (Pair 2). Participants had five seconds to respond before the page automatically moved on to the next token presentation. Following four practice trials, participants moved on to the experimental trials, presented in the same format as the practice, and were given short breaks in between blocks.

3.4. Analysis of production and perception tasks

Time-aligned, word- and phoneme-segmented TextGrid files were generated in *Praat* (Boersma & Weenink, 2019) using the *Montreal Forced Aligner*

(McAuliffe et al., 2017) with a Spanish dictionary (Morgan, 2017), and were subsequently hand-corrected. Three separate acoustic measurements of duration, intensity, and COG were extracted from 220 tokens of intervocalic /b/. Duration was calculated as a relative measurement, by dividing the duration of the /b/ segment by the duration of the vowel-/b/-vowel segment. The right-most boundary of the intervocalic /b/ segment was set using a *Praat* script (Mazzaro, 2011) at the point where the slope of intensity was at its maximum, and the left-most boundary of the segment was set at the point where F3 and F4 began to lower and the periodicity of the waveform changed (Mazzaro, 2011). The second acoustic measurement, relative intensity, was calculated as the difference between the minimum intensity of the /b/ segment and the maximum intensity of the following vowel segment. The resulting intensity values were z-scored to remove talker variability. The last acoustic measurement, COG, was calculated by applying a high-pass filter to the entire sound file and obtaining the measurement from the central 30 ms of each /b/ segment (Mazzaro, 2011), which again was z-scored to remove speaker variability.

As a forced-choice task, the 4I2AFC discrimination task is considered a low-bias task, compared with other discrimination tasks (Wickens, 2001). Therefore, the measure of d-prime submitted to statistical analysis was based on sensitivity alone and was equated to the percent accuracy achieved. Accordingly, the raw perception data from the discrimination task in Qualtrics was exported and coded in the following manner: “1” was assigned for all trials in which the perceived pair with the difference corresponded to the pair with the acoustic difference; “0” was assigned for all trials in which the perceived pair with the difference did not correspond to the acoustically different pair, or otherwise no answer was given in the allotted time.

3.5. Regression models

To analyze the production data, measures of relative duration, z-scored center of gravity, and z-scored relative intensity were submitted as dependent variables to separate linear mixed effects regression models in *R* (R Core Team, 2018) with *lmerTest* (Kuznetsova et al., 2019). Each regression model included main effects of LANGUAGE PROFILE (early, late) and ORTHOGRAPHY (, <v>) and a random intercept of PARTICIPANT. For these regression models, and all subsequent models, the *emmeans* package (Lenth, 2021) was used to calculate Cohen's *d* effect sizes for pairwise comparisons and to perform necessary post-hoc tests using a Tukey pairwise comparison. The *heplots* package (Fox et al., 2021) was used to calculate partial eta-squared (η^2_p) effect sizes for fixed effects models and the *r2glmm* package (Jaeger, 2017) was used to calculate marginal R-squared (R^2) effect sizes for mixed effects models.

To analyze the perception data, the average of each acoustic metric and the average perceptual accuracy were calculated across each speaker using the *dplyr* package (Wickham et al., 2020) in *R*. In order to evaluate the effect of production upon perceptual accuracy (% accuracy), each acoustic measurement (relative duration, z-scored intensity, and z-scored COG), averaged by speaker, was then submitted to a separate fixed effects linear regression model, built using *lmerTest* (Kuznetsova et al., 2017), as an independent variable interacting with LANGUAGE PROFILE. A separate mixed effects logistic regression model predicting perceptual accuracy was built to contain the fixed effect of ORTHOGRAPHY (or <v>), a two-way interaction between LANGUAGE PROFILE (early or late bilingual) and INTERVAL PAIR (six levels), and a random intercept of PARTICIPANT. Significant results were then plotted with *ggplot2* (Wickham, 2016).

4. Results

4.1. Production

Regression coefficients of the mixed effects linear model (Table 3) predicting relative duration and containing main effects of LANGUAGE PROFILE (early bilingual, late bilingual) and ORTHOGRAPHY (, <v>) suggest that neither LANGUAGE PROFILE nor ORTHOGRAPHY are significant predictors of

relative duration. The regression coefficients for the mixed effects linear model (Table 4) predicting z-score center of gravity indicate similarly, where neither the main effect of LANGUAGE PROFILE nor the main effect of ORTHOGRAPHY is a significant predictor of the acoustic metric. However, the effect of LANGUAGE PROFILE is approaching significance ($\beta = 0.30$, $R^2 = 0.020$, $p < 0.066$) where late bilinguals would produce /b/ with slightly higher center of gravity, indicating a more posterior place of articulation. The mixed effects linear regression model predicting z-scored intensity (Table 5) shows a significant main effect of ORTHOGRAPHY ($\beta = 0.28$, $R^2 = 0.023$, $p < 0.01$), where words with <v> are produced with higher intensity by all participants, suggesting a more constricted production. LANGUAGE PROFILE, however, was not a significant main effect of z-scored intensity.

4.2. Perception

Regression coefficients of the fixed effects linear model (Table 6) predicting perceptual accuracy and containing the interaction of RELATIVE DURATION and LANGUAGE PROFILE (early bilingual, late bilingual) suggest that neither LANGUAGE PROFILE nor RELATIVE DURATION are significant predictors of perceptual accuracy. Table 7 contains the output of the fixed effects linear regression model predicting perceptual accuracy with the interaction term of COG and LANGUAGE PROFILE. In this model, neither LANGUAGE PROFILE nor COG are significantly correlated to accuracy. The output for the last fixed effects linear regression model predicting perceptual accuracy contains the interaction of INTENSITY and LANGUAGE PROFILE indicates that neither variable is a significant predictor of accuracy (Table 8).

To observe a relationship between orthography and perceptual discrimination, and between perceptual discrimination and the interaction between language profile and interval pair, a seventh regression model was created. The regression coefficients (Table 9) indicate that ORTHOGRAPHY is a significant predictor of perceptual discrimination ($\beta = -0.20$, $R^2 = 0.002$, $p < 0.05$), such that when ORTHOGRAPHY changes from to <v>, the perceptual accuracy decreases from 59.5% to 54.7%. When perceptual accuracy is plotted over the factors of interval pair and orthography (Figure 5), the decrease in accuracy with <v> grapheme over every interval pair is further observed. Tukey post-hoc tests revealed that

the interaction between LANGUAGE PROFILE and INTERVAL PAIR is not statistically significant, meaning that perceptual accuracy is not systematically different across interval pairs for the two participant groups, nor does it differ for certain interval pairs. However, the differences among pairs 5-8 and 6-9 indicated in the regression output are reflected when perceptual accuracy is plotted over the factors of interval pair and language profile (Figure 6). The two groups of speakers appear to differ in accuracy for interval pairs 1-4, 2-5, and 6-

9, where early bilinguals appear to have better accuracy among more [β]-like tokens (pairs 1-4 and 2-5) and the late bilinguals appear to have better discrimination between the pair of tokens that are more [v]-like (pair 6-9). Though the net effect of language profile is not significant, and the interaction of language profile and interval pair was not significant in post-hoc tests, the visualization of accuracy across language profile suggests that with higher statistical power, the patterns observed visually may surface statistically.

	Estimate	Std. Error	t value	p value	
(Intercept)	0.184962	0.007976	23.191	<2e-16	***
Late	-0.01570	0.010147	-1.547	0.136	
<v>	0.006467	0.005460	1.184	0.238	

Table 3. Regression coefficients for the mixed effects linear regression model predicting relative duration with main effects of LANGUAGE PROFILE (Early, Late) and ORTHOGRAPHY (, <v>) and random intercept of PARTICIPANT. The model intercept is the relative duration of /b/ with as produced by early bilinguals.

	Estimate	Std. Error	t value	p value	
(Intercept)	-0.19956	0.13205	-1.511	0.1390	
Late	0.30048	0.15554	1.932	0.0664	.
<v>	0.07134	0.13025	0.548	0.5845	

Table 4. Regression coefficients for the mixed effects linear regression model predicting z-score COG with main effects of LANGUAGE PROFILE (Early, Late) and ORTHOGRAPHY (, <v>) and random intercept of PARTICIPANT. The model intercept is the z-score COG of /b/ with as produced by early bilinguals.

	Estimate	Std. Error	t value	p value	
(Intercept)	-0.28257	0.24565	01.150	0.26162	
Late	0.26376	0.32736	0.806	0.42903	
<v>	0.27741	0.08697	3.190	0.00166	**

Table 5. Regression coefficients for the mixed effects linear regression model predicting z-score intensity with main effects of LANGUAGE PROFILE (Early, Late) and ORTHOGRAPHY (, <v>) and random intercept of PARTICIPANT. The model intercept is the z-score intensity of /b/ with as produced by early bilinguals.

	Estimate	Std. Error	t value	p value	
(Intercept)	49.71	19.51	2.548	0.0202	*
Duration	35.66	102.89	0.347	0.7329	
Late	20.84	24.93	0.836	0.4141	
Duration: Late	-121.75	136.10	-0.895	0.3828	

Table 6. Regression coefficients for the fixed effects linear regression model predicting perceptual accuracy with an interaction of RELATIVE DURATION and LANGUAGE PROFILE (Early, Late). The model intercept is the accuracy of early bilinguals.

	Estimate	Std. Error	t value	p value	
(Intercept)	54.899	2.726	20.141	8.53e-14	***
COG	-9.238	8.440	-1.095	0.288	
Late	1.097	3.536	0.310	0.760	
COG: Late	7.030	9.846	0.714	0.484	

Table 7. Regression coefficients for the fixed effects linear regression model predicting perceptual accuracy with an interaction of CENTER OF GRAVITY (COG) and LANGUAGE PROFILE (Early, Late). The model intercept is the accuracy of early bilinguals.

	Estimate	Std. Error	t value	p value	
(Intercept)	56.5326	2.4713	22.876	9.36e-15	***
RI	0.8044	3.1453	0.256	0.801	
Late	-0.7775	3.3388	-0.233	0.818	
RI: Late	-1.3094	4.2879	-0.305	0.764	

Table 8. Regression coefficients for the fixed effects linear regression model predicting perceptual accuracy with an interaction of RELATIVE INTENSITY (RI) and LANGUAGE PROFILE (early, late). The model intercept is the accuracy of early bilinguals.

	Estimate	Std. Error	z value	p value	
(Intercept)	0.38475	0.16493	2.333	0.0197	*
<v>	-0.19756	0.07926	-2.493	0.0127	*
Late	-0.35349	0.21592	-1.637	0.1016	
2-5	0.12530	0.20440	0.613	0.5399	
3-6	0.12532	0.20440	0.613	0.5398	
4-7	0.10419	0.20419	0.510	0.6099	
5-8	-0.30600	0.20238	-1.512	0.1305	
6-9	-0.18418	0.20249	-0.910	0.3631	
Late: 2-5	0.06098	0.27513	0.222	0.8246	
Late: 3-6	0.31881	0.27610	1.155	0.2482	
Late: 4-7	0.28763	0.27566	1.043	0.2968	
Late: 5-8	0.59452	0.27391	2.171	0.0300	*
Late: 6-9	0.68105	0.27502	2.476	0.0133	*

Table 9. Regression coefficients for the mixed effects logistic regression model predicting discrimination (0 = “miss”, 1 = “hit”) with a fixed effect of ORTHOGRAPHY, a two-way interaction term of INTERVAL PAIR and LANGUAGE PROFILE, and the random intercept of PARTICIPANT. The intercept is the discrimination of interval pair 1-4 with orthographic by early bilinguals.

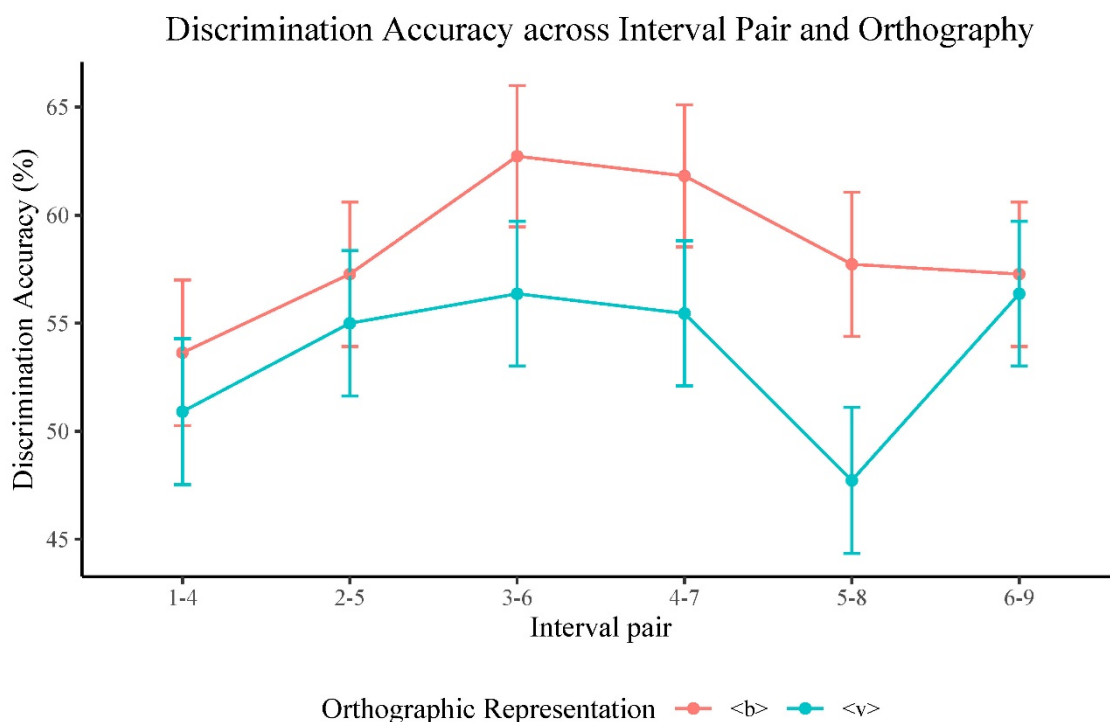


Figure 5. Discrimination accuracy across interval pair and orthography, averaged over all participants.

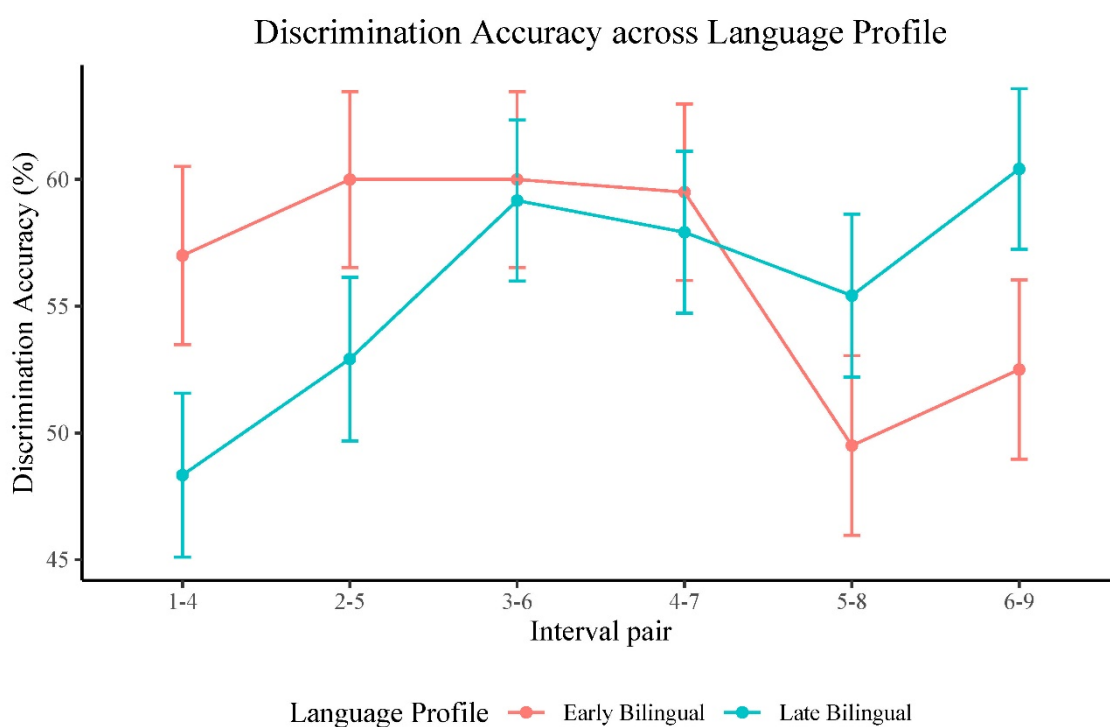


Figure 6. Discrimination accuracy across interval pair and language profile.

5. Discussion and conclusions

This study sought to observe the effects of language profile and orthography on production, as well as the effect of production, language profile, and orthography upon perceptual discrimination of

allophones in a continuum of $[\beta]$ to $[v]$. Having reviewed the results of the experiment in this study, we determine that language profile (i.e., early or late bilingual) is not a significant factor in mediating differences of acoustic measures of intervocalic /b/ production in read speech, although the factor was

approaching significance for z-score COG. However, with higher statistical power, language profile may condition the perception of the most [β]-like and the most [v]-like tokens on the continua. Visually, the early bilinguals appear to have better discrimination accuracy at the more [β]-like end of the continua, and late bilinguals appear to have better discrimination accuracy at the more [v]-like end of the continua. In addition, orthography is a significant predictor of z-score intensity in production, whereby words with grapheme <v> are produced with higher intensity, evidencing a higher degree of constriction. Orthography was also a significant predictor of discrimination accuracy, as the words presented with the grapheme <v> corresponded to lower accuracy.

Due to their earlier and more extensive input from English, participants categorized as early bilinguals were predicted to produce more labiodentalized productions of /b/ (i.e., more [v]-like). However, the data from the present study do not reveal a significant difference in production across early and late bilinguals, regardless of the acoustic metric used to measure production. Participants' productions of /b/, measured by duration, COG, and intensity, were not found to have significant effects on their discrimination accuracy of /b/ allophones. The variability in production, due in part to the allophonic and articulatory similarity of [β] and [v]—with the former having weak contact of the articulators—may explain this. These findings may also be resultant from the way participants were generally grouped, according to age of acquisition of English and not a more holistic measure of language dominance and variety of heritage Spanish. Additionally, the difficulty to discriminate across language profile groups may be exacerbated by the absence of visual cues in this study—i.e. articulatory gestures—that all listeners interacting with /b/ allophones in Spanish perhaps rely on most to distinguish /b/ allophones (as found to be significant in Sadowsky, 2010). Lastly, a study examining whether or not cross-linguistic phonological and orthographic similarity affects cognate word recognition (Carrasco-Ortiz et al., 2021) found that the facilitatory effect of English-language orthography was stronger when Spanish-dominant participants were reading Spanish text. This can be explained by the fact that even Spanish-dominant speakers in the United States often have English-dominant educational experiences and

relatedly, a high or higher reading exposure to the English language rather than Spanish. These findings suggest that phonological similarity benefits from orthographic congruence across different linguistic systems (p. 408), a facilitatory process that all our participants (both early and late bilinguals) utilized in the identification of /b/ allophones in Spanish.

Early bilinguals were predicted to evidence lower discrimination accuracy than late bilinguals, due to their assumed exposure to more variable productions of intervocalic /b/. Language profile was not a significant predictor of perceptual accuracy, indicating that early bilinguals did not have lower accuracy when averaged across all interval pairs. However, the data suggest that with more statistical power, early bilinguals may have greater discrimination accuracy between pairs of tokens that are more [β]-like and late bilinguals may have greater discrimination accuracy between pairs of tokens that are more [v]-like. We interpret these results in terms of the variability of input to which each group may be exposed. The early bilinguals, with more exposure to English [v] and labiodental productions of /b/ in Spanish, may have more difficulty in discriminating tokens within the [v] allophone. On the other hand, late bilinguals have less exposure to labiodental productions and therefore may more easily discriminate between labiodental tokens. Regarding discrimination among pairs of tokens that are more [β]-like, late bilinguals may have more narrowly defined expectations for /b/ and thus more difficulty discriminating between tokens of the same allophonic category. These results will be further analyzed below in a discussion of exemplar-based theories of bilingual phonetic representations.

Considering prior findings of orthography effects on perception, notably that of orthographic and phonological similarity in acquisition observed by Schwartz et al. (2007) in Spanish and Escudero and Wanrooij (2010) in English, the present experiment has evaluated orthography as an indicator of production and perception. Results indicate that orthography was a significant predictor of perceptual accuracy as hypothesized, wherein accuracy decreased when participants were presented with the grapheme <v>. Notably, this effect was observed across both language profiles. As previously mentioned, Archaic Theory posits

that labiodental productions of /b/ in Spanish may be the result of the maintenance of a medieval Spanish phone /v/, which in some varieties of Spanish may not have merged with /b/ (Lope Blanch, 1988; Trovato, 2018). For instance, in Chilean Spanish, production and perception were not found to be dependent on orthography (Sadowsky, 2010; Vergara & Pérez, 2013). Yet, in our production and perception data, orthography was observed to be a significant predictor of relative intensity (i.e., degree of frication) and perceptual accuracy, contradicting the notion that orthographic representation would have no bearing on production or perception of /b/. This supports our hypothesis that in US Spanish, save in varieties of Spanish with historic ties to Peninsular Spanish (e.g., New Mexican Traditional Spanish), labiodental productions may not be attributed to Archaic theory. However, a study that examines these allophones in spontaneous and not read speech may provide different results.

It has also been posited that emphatic speech or hypercorrection, commonly observed in more formal styles and predominantly in Mexican Spanish (Lope Blanch, 1998), may be responsible for more labiodental productions. Due to the frequent correlation between written language use and formal speech styles, orthography may affect how a speaker hypercorrects in formal settings. However, we do not attribute the effects of orthography to hypercorrection. First, the task type utilized in the present experiment did not seek to elicit pedantic speech. Secondly, should labiodental productions be associated with pedantic speech and hypercorrection, discrimination should be more categorical due to the salience of more labiodental variants. However, as interval pair was not a significant predictor of discrimination accuracy, it seems that speakers do not have categorical discrimination of more bilabial and more labiodental variants. Therefore, that orthography is a significant predictor of production and discrimination accuracy supports our initial assumption that labiodental productions in US Spanish are due to language contact with English. This will be further explored below in the discussion of models of bilingual representation.

Though not statistically significant, the fact that interval pair 5-8 yielded a low discrimination accuracy as compared to previous step pairs could

indicate that an allophonic boundary established with acoustic signals occurs somewhere between steps 4 and 5. As steps 5 and 8 are both associated with the same allophonic category and within category discrimination is more difficult than discrimination across category boundaries, overall discrimination accuracy for this interval pair is lower. Interestingly, the difference in accuracy between interval pair 5-8 and 6-9 could be attributed to the fact that the labiodental endpoint (step 9) was unnaturally elicited from our speaker; that is, he does not generally have a more labiodental /b/ in his phonetic repertoire. The artificial nature of this elicited [v] token may cause listeners to perceive step 9 as a third allophonic category, as an artificial [v] distinct from a more naturally produced [v] in Spanish, thereby yielding heightened discrimination accuracy between steps 6 (a more natural labiodentalized token) and 9 (a more artificial labiodentalized token) of the continuum.

Although language profile was not a significant predictor for any of the three acoustic metrics, orthography was a significant predictor of relative intensity, where words with grapheme <v> are correlated with productions of /b/ with higher relative intensity. As relative intensity is a measure of the difference in decibels between a consonant and following vowel, a phone with a greater degree of constriction (i.e., more frication) is expected to have a larger value of relative intensity than a phone with less constriction. That only relative intensity was correlated to orthography suggests that the acoustic difference between bilabial and labiodental tokens resulting from differences in manner of articulation is more prominent than the acoustic difference resulting from differences in place of articulation. Additionally, the lack of significance of both COG and relative duration suggests that these measures are not reliable metrics for distinguishing among allophones of /b/. Our results complement findings in Trovato (2018) where token consonants labeled as labiodental by listeners were more likely to be produced with higher relative intensity (p. 101). Although relative duration was a significant predictor of orthography in Trovato's (2018) study, the effect was strongest when /b/ was word-initial rather than word-medial. Therefore, it is possible that orthography may surface as a significant correlate of relative duration when the context of /b/ is expanded beyond the intervocalic context and includes word-initial contexts. Furthermore, as an

elicitation procedure will often yield more-monitored speech, resulting in lower frequency of less-normative variants, such as a labiodentalized /b/, a production task capable of eliciting less-monitored speech, such as a sociolinguistic interview, may yield more labiodentalized productions of /b/, thereby allowing additional acoustic correlates to surface.

According to the SLM and SLM-r, a speaker acquiring an L2 will form auditory equivalence classes, derived from the statistical properties of the input distributions they have been exposed to in the L2 (Flege & Bohn, 2021). However, as a speaker with a pre-existing L1 phonetic system, the formation of L2 phonetic categories involves disrupting the perceptual links between L1 and L2 categories as phonetic differences between the categories are discerned (Flege & Bohn, 2021, p. 19). Should phonetic differences between categories not be discerned by a speaker, the linked L1 and L2 categories will form a composite category with shared features. In the case of /b/ production in US Spanish, a composite L1-L2 category for /b/ could be associated with a wider range of membership, correlating to more bilabial and more labiodental variants. Though the SLM was formulated around the idea that L2 perception precedes and informs L2 production, the SLM-r clarifies the relationship between production and perception to be bidirectional and potentially asymmetrical. The implication of this directionality is that perception may evidence the distinction of allophones whereas production may not, or vice versa.

A central tenet of the SLM-r is that the objects of perception are position-sensitive allophones, rather than phonemes (Flege & Bohn, 2021, pp. 13-14). Our findings do not directly support this theory, as our stimuli were limited to one phonetic context (i.e., intervocalic /b/). However, our data do suggest that listeners' perceptual sensitivity is context-dependent, where context is equated with the orthographic representation of the lexical item, rather than the identity of surrounding phones. To more fully account for this result, we draw on exemplar models of language representation in addition to the predictions of SLM and SLM-r. In usage-based exemplar models, language use and experience affect cognitive representation and organization of sounds (Bybee, 2002; Johnson, 1997; Pierrehumbert, 2001). Exemplars are stored in

memory as a set of auditory properties and a set of category labels, where the labels may include any classification important to the perceiver (Johnson, 1997, p. 147), such as the gender of the speaker, or in the present study, the orthographic representation of the sound. Exemplars cluster together in clouds based on similarity and the strength of an exemplar is correlated to its frequency of experience (Bybee, 2002). Within this framework, a speaker exposed to bilabial and labiodental variants of Spanish /b/ may have some exemplars associated with the grapheme and some associated with the grapheme <v>, where these exemplars aggregate based on orthographic and auditory similarity. The cluster of exemplars with corresponding grapheme <v> may be represented with more internal variability with respect to acoustic cues of place and manner of articulation, as this context has been seen to yield greater variability in production. Therefore, the larger exemplar cloud corresponding to the grapheme <v> can account for the decreased perceptual discrimination with a stimulus that contains the grapheme <v> as the variability in category membership is greater. In contrast, the grapheme may be linked with a more stringent category with less variability in place and manner of articulation, thereby resulting in better perceptual discrimination when a stimulus contains the grapheme .

The same theoretical perspective can account for the differences that might arise in discrimination accuracy between early and late bilinguals, with the addition of more data. With greater exposure to more [v]-like exemplars of /b/, the exemplar cloud of /b/ of an early bilingual may be larger and may have more internal variability ranging from [β] to [v]. The larger exemplar cloud may yield decreased discrimination accuracy when two exemplars that are [v]-like are both considered to be "in category". Alternatively, a late bilingual may have less exemplars of labiodental /b/ in their exemplar cloud, yielding greater discrimination accuracy. However, the increased density of more [β]-like exemplars in the exemplar cloud of the late bilingual would yield decreased perceptual accuracy when both tokens in a pair are very [β]-like. Accordingly, our findings align with a theoretical framework that combines principles of the SLM-r and exemplar models of speech perception, in which the bilingual representation of sounds from the L1 and L2 are stored in a shared phonetic space and perceptual

discrimination, or lack thereof, is correlated to experience with tokens of sets of auditory properties and category labels, such as orthographic representation.

Lastly, it is vital to consider the social salience of a variable like [v] in US Spanish, where contact with English has created a unique repertoire in part due to phenomena of language contact. While participants in this study and in that of Figueroa Candia and Evans (2021) were not able to categorically discriminate /b/ allophones in experimental tasks, Chappell's (2019, 2020) findings demonstrate that /b/ allophones that are embodied by speakers who present extralinguistic information (i.e., gender and age) are more likely to become salient to listeners. In a matched guise experiment wherein heritage speakers evaluated allophones of /b/ in the speech of early and late bilinguals, Chappell (2019) found that listeners rated late bilinguals and female speakers (of both the late and early bilingual groups) who produced [v] in their Spanish as more intelligent, hard-working, and competent in their Spanish language speaking skills. Male speakers in both groups, on the other hand, were afforded lower ratings in all the same categories, and these evaluations paralleled those of monolingual Mexican speakers (Chappell, 2020). If listeners are aware of the presence of [v] in US Spanish, and [v] indexes different social meanings depending on the speaker, the inclusion of social variables in the token stimuli may change listeners' ability to discriminate /b/ allophones differently. While our study did not consider social factors such as gender and age in the token stimuli nor the participant pool, the inclusion of extralinguistic information is worth exploring to better understand how certain variants are marked, creating positive or negative indexical pathways.

Acknowledgements

The researchers would like to thank: Dr. Keith Johnson (University of California, Berkeley) and Kaile Zhang (The Hong Kong Polytechnic University) for their consultation with respect to research design, stimuli creation, and statistical analysis; Ernesto Gutiérrez-Topete for assisting in the token creation; our colleagues for their valuable input; the UC Berkeley Linguistics Department for funding this project; Damaris and Italo Caro for their participant referrals; our participants for their

cooperation in completing the experiment; and our anonymous reviewers for useful edits and valuable comments. All remaining errors are our own.

References

- Alonso, A. (1967). *De la pronunciación medieval a la moderna en español* (Vol. 1). Gredos.
- Amengual, M. (2019). Type of early bilingualism and its effect on the acoustic realization of allophonic variants: Early sequential and simultaneous bilinguals. *International Journal of Bilingualism*, 23(5), 954–970.
- Bassetti, B., Escudero, P., & Hayes-Harb, R. (2015). Second language phonology at the interface between acoustic and orthographic input. *Applied Psycholinguistics*, 36(1), 1–6.
- Boersma, P., & Weenink, D. (2019). *Praat v.6.0.64*, www.praat.org
- Bybee, J. (2000). Lexicalization of sound change and alternating environments. In M. Broe, & J. Pierrehumbert (Eds.), *Papers in Laboratory Phonology V: Acquisition and the lexicon* (pp. 250–268). Cambridge University Press.
- Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, 14(3), 261–290.
- Canfield, D. L. (1981). *Spanish pronunciation in the Americas*. University of Chicago Press.
- Carrasco, P., Hualde, J. I., & Simonet, M. (2012). Dialectal differences in Spanish voiced obstruent allophony: Costa Rican versus Iberian Spanish. *Phonetica*, 69(3), 149–179.
- Carrasco-Ortiz, H., Amengual, M., & Gries, S. T. (2021). Cross-language effects of phonological and orthographic similarity in cognate word recognition: The role of language dominance. *Linguistic Approaches to Bilingualism*, 11(3), 389–417.
- Chappell, W. (2019). The sociophonetic perception of heritage Spanish speakers in the United States: Reactions to labiodentalized <v> in the speech of late immigrant and US-born voices. In Chappell, W. (Ed.), *Recent advances in the study of Spanish sociophonetic perception* (pp. 240–264). John Benjamins.
- Chappell, W. (2020). Perceptions of Spanish(es) in the United States: Mexicans' sociophonetic evaluations of [v]. In S. Alvord, & G. Thompson (Eds.), *Spanish in the United States* (pp. 31–55). Routledge.

- Colantoni, L., & Marinescu, I. (2010). The scope of stop weakening in Argentine Spanish. In Ortega-Llebaria, M. (Ed.), *Selected proceedings of the 4th Conference on Laboratory Approaches to Spanish Phonology* (pp. 100–114). Cascadilla Press.
- Cole, J., Hualde, J. I., & Iskarous, K. (1999). Effects of prosodic and segmental context on /g/-lenition in Spanish. In O. Fujimura, B. D. Joseph, & B. Palek (Eds.), *Proceedings of the Fourth International Linguistics and Phonetics Conference* (pp. 575-589). Karolinum Press.
- Eddington, D. (2011). What are the contextual phonetic variants of /β, ð, γ/ in colloquial Spanish?. *Probus*, 23(1), 1–19.
- Escudero, P., & Wanrooij, K. (2010). Learning the phonological forms of new words: Effects of orthographic and auditory input. *Language and Speech*, 53, 361–381.
- Figuerola Candia, M. A., & Evans, B. G. (2021). Perception of the bilabial-labiodental contrast in the approximant consonants of the Chilean Spanish. *Loquens*, 7(1), e067–e067.
- Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience. Issues in Cross-language Research*, 92, 233–277.
- Flege, J. E., & Bohn, O. (2021). The Revised Speech Learning Model (SLM-r). In R. Wayland (Ed.), *Second language speech learning: Theoretical and empirical progress* (pp. 3–83). Cambridge University Press.
- Fox, J., Friendly, M., & Monette, G. (2021). *heplots: Visualizing tests in multivariate linear models. R package version 1.3-8*. <https://cran.r-project.org/web/packages/heplots/>
- Gerrits, E., & Schouten, M. E. H. (2004). Categorical perception depends on the discrimination task. *Perception & Psychophysics*, 66(3), 363-376.
- Gertken, L. M., Amengual, M., & Birdsong, D. (2014). Assessing language dominance with the bilingual language profile. In P. Leclercq, A. Edmonds, & H. Hilton (Eds.), *Measuring L2 proficiency: Perspectives from SLA* (pp. 208–226). Multilingual Matters.
- Guasch, M., Boada, R., Ferré, P., & Sánchez-Casas, R. (2013). NIM: A Web-based Swiss Army knife to select stimuli for psycholinguistic studies. *Behavior Research Methods*, 45, 765-771.
- Han, J. I., & Choi, T. H. (2016). The influence of spelling on the production and storage of words with allophonic variants of /h/ in Korean, *Applied Psycholinguistics*, 37(4), 757-780.
- Hualde, J. I. (2013). *Los sonidos del español*. Cambridge University Press.
- Hyltenstam, K., Bylund, E., Abrahamsson, N., & Park, H.-S. (2009). Dominant-language replacement: The case of international adoptees. *Bilingualism: Language and Cognition*, 12(2), 121–140.
- Jaeger, B. (2017). *r2glmm: Computes R squared for mixed (multilevel) models*. R package version 0.1.2. <https://cran.r-project.org/web/packages/r2glmm/index.html>
- Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson, & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–166). Morgan Kaufmann.
- Kataoka, R., & Johnson, K. (2007). Frequency effects in cross-linguistic stop place perception: A case of /t-/k/ in Japanese and English. *UC Berkeley PhonLab Annual Report*, 3(3), 273–301.
- Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE (pp. 3933-3936). Institute of Electrical and Electronics Engineers.
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). *lmerTest package: Tests in linear mixed effects models*. *Journal of Statistical Software*, 82(13), 1-26.
- Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages* (Vol. 1012). Blackwell.
- Ladefoged, P. (2003). *Phonetic data analysis: An introduction to fieldwork and instrumental techniques*. Wiley-Blackwell.
- Lenth, R. V. (2021). *emmeans: Estimated marginal means, aka least-squares means*. R package version 1.6.0. <https://cran.r-project.org/web/packages/emmeans/index.html>

- Lipski, J. M. (1994). Spanish stops, spirants and glides: From consonantal to [vocalic]. In M. Mazzola (Ed.), *Issues and theory in Romance Linguistics (selected papers from the Linguistic Symposium on Romance Languages XXIII)* (pp. 67-86). Georgetown University Press.
- Lope Blanch, J. M. (1988). La labiodental sonora en el español de México. *Nueva Revista de Filología Hispánica*, 36(1), 153-170.
- Martínez-Gil, F. (1998). On the spelling distinction b vs. u/v and the status of spirantization in Old Spanish. In J. Lema, & E. Treviño (Eds.), *Theoretical analyses on Romance Languages: Selected papers from the 26th Linguistic Symposium on Romance Languages* (pp. 283-316).
- Mayer, J. (2013). *Concatenate with pause*. Praat script. <http://praatpfaane.lingphon.net/praat-ressourcen/resources-english>.
- Mazzaro, N. (2011). *Experimental approaches to sound variation: A sociophonetic study of labial and velar fricatives and approximants in Argentine Spanish* (PhD dissertation). University of Toronto.
- Mazzaro, N., & González de Anda, R. (2019). The perception-production connection: /tʃ/ deaffrication and rhotic assibilation in Chihuahua Spanish. In Chappell, W. (Ed.), *Recent advances in the study of Spanish sociophonetic perception* (pp. 288-311). John Benjamins.
- McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. & Sonderegger, M. (2017). *Montreal Forced Aligner*. <https://montrealcorpus-tools.github.io/Montreal-Forced-Aligner/>
- Morgan, J. (2017). *Santiago Spanish Lexicon: A pronouncing dictionary for the Spanish language*, <https://www.openslr.org/34/>
- Ortega, R. I. (2018). *A voiced labiodental fricative [v] in El Paso Spanish* (Master's thesis). University of Texas at El Paso.
- Ortega-Llebaria, M. (2003). Effects of phonetic and inventory constraints in the spirantization of intervocalic voiced stops: Comparing two different measurements of energy change, *Stress*, 7(6), 2817-2820.
- Penny, R. (2002). *A history of the Spanish language*. Cambridge University Press.
- Pierce, L. J., Klein, D., Chen, J. K., Delcenserie, A., & Genesee, F. (2014). Mapping the unconscious maintenance of a lost first language. *Proceedings of the National Academy of Sciences*, 111(48), 17314-17319.
- Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee, & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 137-157). John Benjamins.
- Qualtrics Labs, Inc. (2019). *Qualtrics*. www.qualtrics.com
- R Core Team (2018). *R: A language and environment for statistical computing v.3.5.2*, <https://www.R-project.org/>
- Rafat, Y. (2015). The interaction of acoustic and orthographic input in the acquisition of Spanish assibilated/fricative rhotics. *Applied Psycholinguistics*, 36(1), 43-66.
- Rastle, K., McCormick, S. F., Bayliss, L., & Davis, C. J. (2011). Orthography influences the perception and production of speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(6), 1588-1594.
- Rao, R. (2014). On the status of the phoneme /b/ in heritage speakers of Spanish. *Sintagma*, 26, 37-54.
- Roelofs, A. (2006). The influence of spelling on phonological encoding in word reading, object naming, and word generation. *Psychonomic Bulletin & Review*, 13(1), 33-37.
- Sadowsky, S. (2010). El alófono labiodental sonoro [v] del fonema /b/ en el castellano de Concepción (Chile): Una investigación exploratoria. *Estudios de Fonética Experimental*, XIX, 231-261.
- Schwartz, A. I., Kroll, J. F., & Diaz, M. (2007). Reading words in Spanish and English: Mapping orthography to phonology in two languages. *Language and Cognitive Processes*, 22(1), 106-129.
- Sebastián-Gallés, N., Echeverría, S., & Bosch, L. (2005). The influence of initial exposure on lexical representation: Comparing early and simultaneous bilinguals. *Journal of Memory and Language*, 52(2), 240-255.
- Stevens, P. (1960). Spectra of fricative noise in human speech. *Language and Speech*, 3, 32-39.
- Temkin Martínez, M. T., & Müllner, I. (2016). Specific exceptions driving variation: The role of orthography in Modern Hebrew spirantization. In *Proceedings of the Annual Meetings on Phonology* (Vol. 3). <https://doi.org/10.3765/amp.v3i0.3695>

- Torres Cacoullos, R., & Ferreira, F. (2000). Lexical frequency and voiced labiodental-bilabial variation in New Mexican Spanish. *Southwest Journal of Linguistics*, 19(2), 1-17.
- Trovato, A. M. (2018). *A sociophonetic analysis of contact Spanish in the United States: labiodentalization and labial consonant variation* (PhD dissertation). University of Texas at Austin.
- Vergara, V., & Pérez, H. E. (2013). Estudio de la incidencia de la representación gráfica (escritura) en la producción del alófono labiodental [v] del fonema /b/ [The role of orthographic representation in the frequency of production of the voiced labiodental fricative [v] allophone of /b/]. *Boletín de Filología*, 48(2), 119–128.
- Wickens, T. (2001). *Elementary signal detection theory*. Oxford University Press.
- Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.
- Wickham, H., François, R., Henry, L. & Müller, K. (2020). *dplyr: A grammar of data manipulation*. R package v.0.5.8. <https://CRAN.R-project.org/package=dplyr>.
- Zhang, Q. & Damian, M. F. (2012). Effects of orthography on speech production in Chinese. *Journal of Psycholinguistic Research*, 41(4), 267-283.

