Dominance, mode, and individual variation in bilingual speech production and perception

Page Piccinini[1] & Amalia Arvaniti[2]

[1]Département d'Études Cognitives, École Normale Supérieure, France

[2]English Language and Linguistics, University of Kent, UK

Abstract:

Early Spanish-English bilinguals and English controls were tested on the production and perception of negative, short-lag, and long-lag *Voice Onset Time* (VOT). These VOT types span Spanish and English phonetic categories. Phonologically, negative and short-lag VOT stops are distinct phonemes in Spanish, while both are realizations of *voiced* stops in English. Dominance was critical: more English-dominant bilinguals produced more short-lag VOT stops in response to negative VOT stimuli, and were less accurate than more balanced bilinguals at discriminating negative from short-lag VOT. Bilinguals performed similarly to monolinguals overall, but they produced more negative VOT tokens and shorter short-lag VOT in response to negative VOT. Their productions were also less well correlated with perception and showed more variation between individuals. These results highlight the variable nature of bilingual production and perception, and demonstrate the need to consider language dominance, individual variation, as well as modalities and tasks when studying bilinguals.

1. Introduction

Since Green (1998), it has been well established that bilingual production and perception are based on a system of activation and suppression of linguistic subsystems, supporting the view that a bilingual's languages inevitably influence each other (cf. Flege, 1995; Grosjean, 1989). However, it is not clear how this model applies to speech production and perception. While Sundara, Polka and Baum (2006) found that bilinguals do not differ from monolinguals in production (contra the above assertion), most studies indicate that each language affects the other, with the magnitude and direction of the effects dependent on task and the type of bilingualism tested. Flege and Eefting (1987) found that the production of English stops by English-Spanish bilinguals is affected by age of acquisition: early bilinguals produced values close to those of monolinguals, but late bilinguals showed influence of L1 on L2. Fowler, Sramko, Ostry, Rowland, and Hallé (2008) reported similar results for English-French bilinguals. Mack (1989) found that while L2-dominant early sequential bilinguals do not differ from monolinguals in production and discrimination, they respond differently from monolinguals in identification tasks. Further, studies on code-switching often show asymmetrical effects on speech production with the non-dominant language affecting the dominant (Flege, MacKay, & Piske, 2002; Olson, 2013; Balukas & Koops, 2014; Piccinini & Arvaniti, 2015).

A possible reason for these diverse results is the reliance on different tasks and on "bilingual" populations that vary substantially in age and manner of acquisition, and in exposure to and use of each language. As Dunn and Fox Tree (2009: 273) note, "[w]hen considering bilingualism globally, perhaps the only thing to be counted on is a diversity of experiences." In the studies reviewed by Dunn and Fox Tree (2009), participants ranged from speakers who grew up monolingual and had only formal instruction in their L2, to speakers who belonged to a bilingual community, used both languages daily, and routinely code-switched. A consequence of this diversity is uneven input and use, leading to differences in language proficiency and dominance, factors that affect bilingual performance and consequently experimental results (Flege et al., 2002). Here we are interested in the bilinguals that Dunn and Fox Tree (2009) term *simultaneous bilingual speakers*, speakers exposed to both languages from a very early age, sometimes from the start of linguistic input.

A corollary of this variety of study conditions is that the relationship between production and perception among bilinguals remains unclear. Most research suggests a close link, with both languages active in both production and perception (Flege, 1995; Green, 1998). Others have argued that production is modelled on each language, but perception is based on the dominant language (Cutler, Mehler, Norris, & Segui, 1992). Still others find that the link is task-dependent: Beach, Burnham, and Kitamura (2001) tested Greek-English bilinguals in an imitation and a discrimination task involving Thai contrasts, and found no correlation between tasks; however, the discrimination results correlated with the production of native contrasts: bilinguals who most differentiated their productions in English and Greek were better at discriminating Thai contrasts.

An additional complication is that research on bilingual speech production and perception has largely focused on *phonetic* categories. One dimension frequently tested is VOT (voice onset time), the interval between the release of a stop closure and the beginning of voicing, perceived as aspiration. VOT forms a continuum from negative values (voicing starts before the stop closure is released) to positive values (voicing starts after closure release). Three phonetic categories are assumed within this continuum: 1) negative VOT (prevoicing), 2) short-lag VOT, where voicing resumes shortly after release, leading to unaspirated stops, and 3) long-lag VOT, where voicing resumes tens to hundreds of milliseconds after release, leading to aspirated stops (Cho & Ladefoged, 1999). Short-lag and long-lag VOT (e.g., [p] and [p^h] respectively) have been extensively investigated in bilingual speech, as they represent typical realizations of phonemes that in many language pairs would lead to "equivalence classification" (Flege, 1995). For example, [p] is the prototypical realization of the phoneme /p/ in Spanish, French, and Greek, while [p^h] is the prototypical realization of /p/ in English. Research has focused on whether under such circumstances "equivalence classification" occurs, with the speaker adjusting her production so VOT is somewhat long for [p] and somewhat short for [p^h], or whether the categories remain distinct (cf. MacKay, Flege, Piske, & Schirru, 2001; Fowler et al., 2008; Grijalva, Piccinini, & Arvaniti, 2013).

Less attention has been paid to the situation that arises when the same phonetic category must be classified differently at the phonological level in the bilingual's two languages. This question is of interest for two reasons: first, it tests whether "equivalence classification" can be obtained in these circumstances; second, it tests whether all subsystems of a bilinguals' two languages are always active (Green, 1998). If so, then the question that arises is which subsystem bilinguals use during speech production and perception to successfully produce and categorize incoming speech segments (Beddor, 2017). With respect to perception in particular, if bilinguals rely on phonological categories to classify incoming segments, then bilinguals of certain languages are faced with the fact that the same segment must be phonologically classified differently in each of their languages. When this happens, how do bilinguals cope? Would they approach the task in a monolingual *set* (Elman, Diehl, & Buchwald, 1977) or *mode* (Grosjean, 2001) based on ambient language, or would language dominance determine performance (cf. Cutler et al., 1992; Flege et al., 2002)? According to Flege (1995: 242) a bilingual's two languages exist in "*a common phonological space*" [emphasis in the original]. If so, how would conflicting phonological classifications operate? Flege (1995) suggests that production categories would become maximally distinct, but his own example of a French-English bilingual child who produced both English and French voiceless stops with long-lag VOT does not support this outcome (Flege, 1995).

To address the matter of categorization, the present study examined early Spanish-English bilinguals' and English monolinguals' abilities to produce and perceive distinctions within the VOT continuum. Although both Spanish and English are phonologically analyzed as contrasting voiced and voiceless stops (/p/~/b/, /t/~/d/, /k/~/g/), these abstract (phonological) categories map onto different phonetic categories in each language, a difference captured by VOT.

Spanish voiced stops show prevoicing (negative VOT), ranging from -77 ms to -100 ms (Flege & Eefting 1987; Lisker & Abramson, 1964; Dmitrieva, Llanos, Shultz, & Francis, 2015). Spanish voiceless stops are unaspirated; they have short-lag VOT, ranging from 4 to 39 ms, depending on rate and place of articulation and dialect (Lisker & Abramson, 1964; Magloire & Green, 1999; Dmitrieva et al., 2015).

English voiceless stops are typically aspirated (long-lag VOT), ranging from 20 to 130 ms depending on place of articulation, dialect, speaking rate, and gender (Lisker & Abramson, 1964; Docherty, 1992; Kessinger & Blumstein, 1997; Magloire & Green, 1999; Dmitrieva et al., 2015); VOT also shows extensive inter-speaker variability (Theodore, Miller, & DeSteno, 2009). Phonologically voiced stops are overwhelmingly realized as voiceless unaspirated stops (Davidson, 2016; Abramson & Whalen, 2017, and references therein). Prevoicing is rare in running speech: Davidson (2016) reports that less than 4% of the (phonologically) voiced stops in her American English corpus were produced with non-residual voicing; similar results are reported by Stuart-Smith, Sonderegger, Rathcke, and Macdonald (2015), on Scottish English, and Nakai and Scobbie (2016) on a variety of English dialects. Prevoicing is more prevalent in citation forms, like those used in the present study, but still infrequent, found in 24-31% of tokens; it is of similar duration to prevoicing in Spanish (MacKay et al., 2001; Dmitrieva et al., 2015).

These differences in realization between English and Spanish stops mean that *phonetically* Spanish-English bilinguals have at their disposal the entire VOT range, but mapped differently onto phonological categories in their two languages. This applies particularly to stops with short-lag VOT which are phonetic realizations of phonologically voiced stops in English, but of voiceless stops in Spanish. This is schematically represented in Figure 1. The focus of the present paper is this specific conflict in phonological categorization.

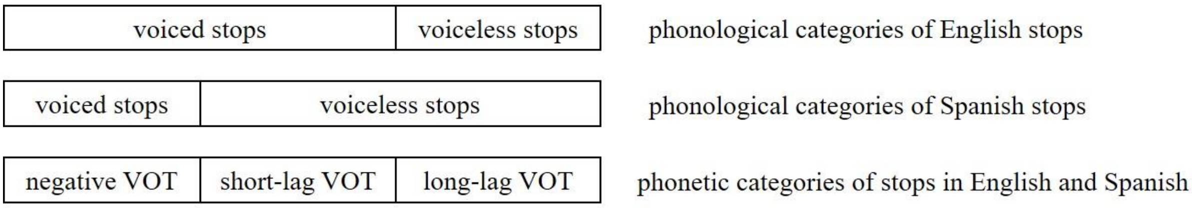
voiced stops
voiceless stops
phonological categories of English stops
voiced stops
voiceless stops
phonological categories of Spanish stops
negative VOT
short-lag VOT
long-lag VOT
phonetic categories of stops in English and Spanish

Figure 1. Schematic representation of Spanish and English VOT categories in phonetics and phonology.

To sum up, this conflict poses the following questions: do bilinguals handle the three VOT categories phonetically or phonologically, or is phonetic detail always available during speech perception? Would language mode or dominance affect performance?

To answer these questions, Spanish-English bilinguals and monolingual English controls were tested on Eastern Armenian, which has a phonological three-way VOT contrast, negative, short-lag, and long-lag VOT; e.g., [bɑh] 'spade' vs. [pɑh] 'movement'; [bɑk] 'courtyard' vs.

[p^{h}ɑk] 'closed, shut' (Dum-Tragut, 2009). First, they participated in a production task in which they heard and had to repeat Armenian words. If participants treated this as an imitation task, they would use the phonetic categories at their disposal to faithfully (i.e., *phonetically*) reproduce the stimuli (cf. Flege & Eefting, 1988). If, however, they categorized the tokens *phonologically* and used prototypical exemplars to reproduce the stimuli, then stimuli with short-lag VOT would pose a problem since they belong to different phonological categories in English and Spanish. In this instance, language dominance or mode could determine which language would be used in perception and production: if English dominated, productions should show fewer tokens with negative VOT in response to negative VOT stimuli, since such productions are rare in American English, as discussed above.

In addition, half of the participants in the production study took part in an AX task, while the other half participated in an ABX task. AX allows listeners to use *auditory mode* (Pisoni, 1973), in other words, to focus on phonetic detail. ABX requires that listeners store stimuli A and B in short-term memory to compare them to stimulus X; this procedure is assumed to require a level of abstraction akin to phonological categorization (Raphael, Borden, & Harris, 2007; McGuire, 2010). Thus, the hypotheses were as follows: if participants have a pool of *phonetic* categories based on their two systems, the three-way contrast of Eastern Armenian should be unproblematic for tasks requiring phonetic categorization (here, the AX task, and the production task if treated as an imitation task). For the ABX task, however, *phonological* categorization was expected. If bilinguals have a common phonological space, as argued by Flege (1995), tokens that have mutually exclusive classifications in the two languages (here, short-lag VOT stimuli) should lead to poor discrimination, as participants would have to simultaneously classify them as voiceless, as in Spanish, *and* as voiced, as in English. Alternatively, if phonological categorization is specific to each language (as argued, e.g., by Cutler et al., 1992) then bilinguals should respond like monolinguals of one or the other of their languages, with mode and dominance influencing which system is preferentially activated.

2. Experiment 1: Production

2.1. Method

2.1.1. Participants

Forty early Spanish-English bilinguals of Mexican-American heritage (30 female), and 40 monolingual speakers of American English (28 female) took part. The bilingual participants had been exposed to both Spanish and English before age six, and fell in the group Dunn and Fox Tree (2009: 273) describe as *simultaneous bilinguals*, "second-generation bilingual language learners within an immigrant family." All participants were undergraduates at UC San Diego and took part in the study in exchange for course credit. As the experiments took place in the USA, it was not possible to recruit 40 comparable, monolinguals speaking the same variety of Spanish as the bilingual group (Border Spanish as spoken in the US-Mexico border along San Diego and Tijuana; [Bills, Chávez, & Hudson, 1995; Lipski, 2008]).[1]

The bilingual participants completed the Bilingual Dominance Scale (henceforth BDS; Dunn & Fox Tree, 2009). The BDS, a scale developed specifically for bilinguals with similar characteristics to the present participants, assigns each speaker a dominance score. While a single number cannot provide a complete picture of a phenomenon as complex as dominance

[1] There is little research on VOT perception by Spanish monolinguals. Keating (1984) suggests they would not have difficulty distinguishing three VOT categories in phonetic tasks. Though the exact category boundaries may differ between monolingual English and Spanish speakers (Williams, 1977), the differences are unlikely to be of relevance to the present study.

(Flege et al., 2002), the BDS is useful for incorporating dominance into statistical modelling. In the BDS, dominance is determined using twelve questions, covering ages of acquisition and proficiency, years of education, current use and preference, current accent, and current country of residence. The scale thus incorporates past research on the factors contributing to dominance (e.g., Grosjean, 1998, on language restructuring; Flege et al., 2002, on age of acquisition and exposure). A score of 0 means the speaker is perfectly balanced; high positive scores indicate heavy dominance in one language, and high negative scores heavy dominance in the other. The present participants had a mean score of 9.23 (median = 9), thus they tended to be English-dominant. The standard deviation was high (8.63) because of two participants with high negative scores (-17 and -12); without them the distribution was normal (Figure 2).

None of the monolingual participants reported significant exposure to any language besides English before the age of 6. Half of the monolinguals had studied through formal instruction a second language that had a contrast between negative and short-lag VOT stops, such as Spanish or French.

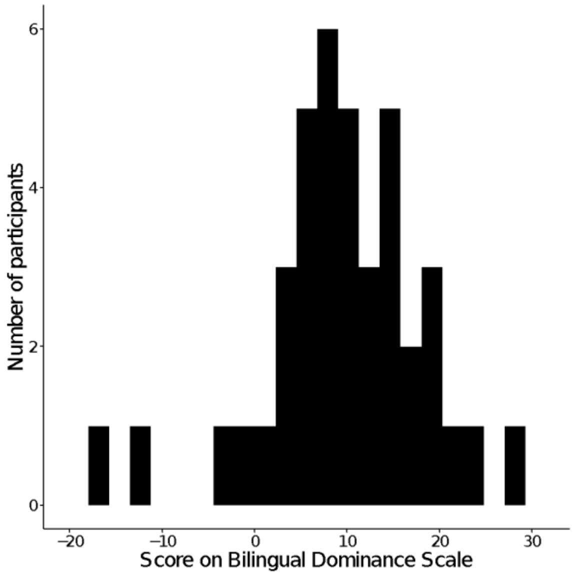

Figure 2. Distribution of bilingual participant scores on the Bilingual Dominance Scale (Dunn & Fox Tree, 2009).

2.1.2. Stimuli

The stimuli were Eastern Armenian words, selected in consultation with a native speaker and supplemented by nonce words and fillers. The test words started with a stop with negative, short-lag or long- VOT in one of two places of articulation, bilabial and velar, giving six phonetic categories [p], [p^h], [b], [k], [k^h], [g]. Coronals were not included as they are alveolar in English but dental in Spanish; this difference could affect VOT duration (Lisker & Abramson, 1964; Cho & Ladefoged, 1999), complicating the comparison of monolingual and bilingual productions; further, the stimuli could sound more English- than Spanish-like, since Eastern Armenian has alveolar stops. Test words and fillers (which begun with [m], [n], [f], or [s]) were divided between monosyllabic C_1VC_2 and disyllabic ˈ$C_1V_1.C_2V_2$ so that phonotactics did not affect listener responses: C_1VC_2 conforms to English phonotactics and ˈ$C_1V_1C_2V_2$ to Spanish. In all words, the initial consonant, which was one of the six stops under investigation, was followed by one of [i], [ɛ], or [o]; [i] is present in all three languages, [ɛ] in Eastern Armenian and English (e.g., in English ‘bed’), and [o] in Eastern Armenian and Spanish. For all words, C_2 was [m]; $C_1V_1C_2V_2$ words ended in either [a] or [o].

Two female heritage speakers of Eastern Armenian, who were UC San Diego students at the time, produced a total of 90 words (10 initial consonants × 3 medial vowels × 3 endings). Two speakers were recorded to ensure greater variability and encourage participants to extract both language-specific and speaker-specific categories from the stimuli; they were also needed for the perception experiments to avoid listeners expecting identical tokens in any given trial. The speakers produced four tokens of each word in isolation. Words were presented on a computer screen in Armenian alphabet and a Latin alphabet based transliteration used by the Armenian diaspora. Before the recording, the speakers went over the materials with the consultant, to ensure they could produce all words correctly. They were instructed to say the nonce words as if they were real words of Eastern Armenian. Some nonce combinations were excluded before the recording because they are real words in English or Spanish (e.g., *beam* [bim], *quemo* [kɛmo] 'I burn'), resulting in a final set of 48 stimuli, 24 test items and 24 fillers. Twenty-two of these were CVC words (12 test items, 10 fillers) and 26 CVCV words (12 test items, 14 fillers).

For the test words, VOT was measured to ensure the three-way contrast was present: [b] = -91.84 ms (*SD* = 31.33), [p] = 22.96 ms (3.59), [p^h] = 93.47 ms (22.52), [g] = -95.05 ms (25.02), [k] = 43.82 ms (8.55), [k^h] = 109.23 ms (25.08). For 48 tokens without appropriate VOT duration, the VOT from a word with the same initial CV sequence replaced the original VOT by splicing at zero-crossings at points of rising amplitude. Within each speaker's data, stimuli were chosen so that the VOT distributions of the three stops (e.g., [b], [p], [p^h]) did not overlap. Unpaired t-tests showed that VOT distributions did not differ significantly between speakers, except for [p^h] and [k^h] for which one speaker had longer VOT.

2.1.3. Procedure

The experiment took place in the Speech Lab of UC San Diego. Before the experiment started, participants were administered a language questionnaire, which for the bilinguals included the BDS (see section 2.1.1). To test whether language mode would affect performance (Grosjean, 2001), half of the bilinguals were tested in English and the other half in Spanish; the language of the questionnaire matched that of the experiment. The two groups were matched for dominance: the English-mode group had an average BDS score of 9.85 (*SD* = 8.37), and the Spanish-mode group an average of 8.60 (*SD* = 9.05) (t = 0.45, p = 0.65). The monolingual participants were tested in English.

The stimuli were presented over headphones using SuperLab Pro 4.5 (Cedrus Corporation, 2011). Participants were told they would hear words in a new language; their task wasto repeat them as best they could. They were instructed to focus on a fixation cross which disappeared immediately after the word finished playing. Participants had to repeat the word as soon as the cross disappeared (i.e., there was minimal lag between hearing a word and producing it). Participants then pressed the spacebar to move to the next trial. The fixation cross reappeared and after 500 ms the next word would play. The stimuli were presented twice across two blocks (and randomized within each block) for a total of 384 productions per participant; 192 of these were test items, evenly divided between bilabials and velars, for a total of 15,360 tokens (192 tokens × 80 participants).

2.1.4. Annotation and measurements

Positive VOT was measured from the release of the stop to the onset of voicing. For negative VOT, measurements were taken from the onset of voicing to the release of the stop and marked as negative; this was the one pattern of prevoicing found in our data (cf. Davidson, 2016). All words were categorically coded for whether the stimulus heard had negative, short-lag, or long-lag VOT (e.g., [b], [p], or [p^h] respectively).

2.1.5. Analyses

Two separate sets of analyses were run, one on bilinguals to examine language mode and dominance, and one comparing monolinguals to bilinguals. All analyses were conducted on the responses to negative and short-lag stimuli, as examination of the data showed that no participants had difficulty producing stops with long-lag VOT. Two analyses were run on the stimuli with negative and short-lag VOT. For the first, each token was coded as being produced with negative VOT (0) or positive VOT (1). The expectation was that tokens produced in response to stimuli with short-lag VOT would have positive VOT more often than tokens produced in response to stimuli with negative VOT. The second analyses focused on short-lag tokens, to test whether they had longer VOT when produced in response to stimuli with short-lag or negative VOT. We focused on productions with positive VOT to ensure the data did not have a bimodal distribution. Separate models were run for bilabial and velar stops due to known differences in VOT duration by place of articulation (Cho & Ladefoged, 1999).

For the analysis of bilinguals, a generalized linear mixed effects model and a linear mixed effects model were run, with production of positive VOT (0, 1) and VOT duration in ms as the dependent variables respectively. For both models, stimulus category (negative, short-lag), language mode (English, Spanish), and BDS score were included as fixed effects together with two interactions, stimulus category × language mode and stimulus category × BDS score. Stimulus category and language mode were contrast coded; BDS score was coded as a numeric variable. For the generalized model on negative vs. positive VOT, participant was included as a random intercept. For the model on positive VOT durations, participant was included as a random intercept and a random slope by stimulus category; word root (e.g., "-ima") was included as a random intercept and a random slope by stimulus category, language mode, and BDS score uncorrelated with the random intercept.

For the analysis of bilinguals vs. monolinguals, similar models were run, the main difference being that the models included stimulus category and language background (monolingual, bilingual) and their interaction as fixed effects. For the generalized model on negative vs. positive VOT, participant was included as a random intercept and a random slope by stimulus category uncorrelated with the random intercept; word root was included as a random intercept. For the model on positive VOT durations, participant was included as a random intercept and a random slope by stimulus category; word root was included as a random intercept and a random slope by the interaction of stimulus category and language background. Stimulus-speaker was not included as a random effect as not all models would converge when it was included. For all models, these were the maximal random effects structures that would converge (Barr, Levy, Scheepers, & Tily, 2013). Significance testing was done with model comparison with alpha set to 0.05.

2.2. Results

2.2.1. Bilinguals

As indicated in Figure 3, bilinguals produced distinct distributions in response to the stimuli with short-lag and long-lag VOT (e.g., [p] and [p^h] respectively); for both categories, the distributions were unimodal. However, their productions in response to stimuli with negative VOT had bimodal distributions for both /b/ and /g/; separate analyses for these are discussed in section 2.1.5.

The comparison of negative vs. positive VOT in bilabials showed a significant effect of stimulus category, such that tokens produced in response to [p] were more likely to be produced with positive VOT than tokens produced in response to [b] ($\beta = 2.41$, $SE = 0.19$, $\chi^2(1) = 227.79$, $p < 0.001$). The interaction of stimulus category and BDS score was also significant ($\beta = -0.07$, *SE* = 0.01, $\chi^2(1) = 26.43$, $p < 0.001$). To better understand the interaction, follow-up simple logistic regressions were run on each stimulus category separately, with negative or positive VOT as the dependent variable and BDS score as the independent variable. As illustrated in Figure 4(a), for the model on the tokens produced in response to stimuli with negative VOT, BDS score was significant, with a higher BDS score (more English-dominant) resulting in more tokens with *positive* VOT ($\beta = 0.06$, $SE = 0.01$, $z = 7.71$, $p < 0.001$); for the model on the tokens produced in response to stimuli with short-lag VOT, BDS score was not significant.

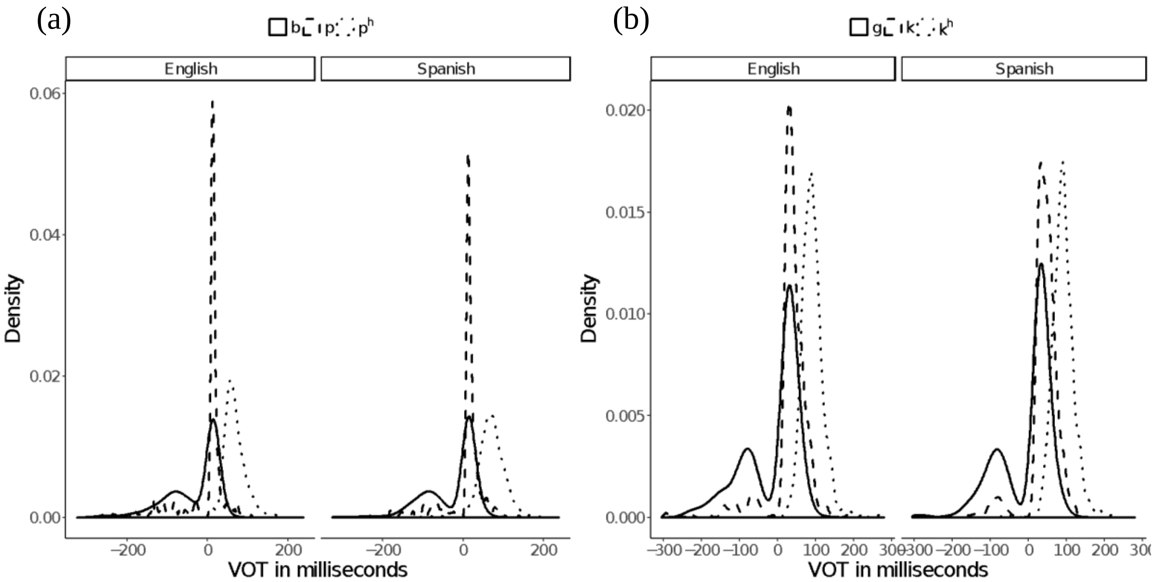

Figure 3. Distributions of bilinguals' VOT values in the production task by language mode and stimulus category for (a) bilabial stops and (b) velar stops.

The duration analysis of positive VOT productions showed a significant effect of stimulus category, such that VOT produced in response to [p] stimuli was longer than positive VOT produced in response to [b] stimuli ($\beta = 10.55$, $SE = 2.17$, $\chi^2(1) = 16.38$, $p < 0.001$). The interaction of stimulus category and BDS score was also significant ($\beta = -0.53$, $SE = 0.17$, $\chi^2(1) = 8.78$, $p = 0.003$). Follow-up simple linear regressions were run on each stimulus category separately, with duration as the dependent variable and BDS score as the independent variable. For the model on the tokens produced in response to stimuli with negative VOT, BDS score was significant, with a higher BDS score resulting in longer VOT durations ($\beta = 0.20$, $SE = 0.05$, $t = 4.14$, $p < 0.001$). For the model on the tokens produced in response to stimuli with short-lag VOT, BDS score was also significant, but with higher BDS score resulting in shorter VOT ($\beta = -0.33$, $SE = 0.06$, $z = -5.52$, $p < 0.001$). Figure 4(b) illustrates this interaction which resulted in more English-dominant participants using similar VOT durations for tokens produced in response to both [p] and [b] stimuli, while those who were more Spanish-dominant, produced shorter VOT in response to [b] relative to [p].

The analysis of negative vs. positive VOT in velars showed a significant effect of stimulus category, such that tokens produced in response to stimuli with short-lag VOT were more likely to have positive VOT than tokens produced in response to stimuli with negative VOT ($\beta = 2.59$, $SE = 0.18$, $\chi^2(1) = 267.22$, $p < 0.001$). The interaction of stimulus category and BDS score was also significant ($\beta = -0.03$, $SE = 0.01$, $\chi^2(1) = 5.67$, $p = 0.02$); see Figure 5. The same follow-up logistic regressions were run as for bilabials. BDS score was significant for both the model on responses to the [g] stimuli and the model on responses to the [k] stimuli, with a higher BDS score resulting in more tokens with positive VOT (for [g]: $\beta = 0.06$, $SE = 0.008$, $z = 8.54$, $p < 0.001$; for [k]: $\beta = 0.04$, $SE = 0.01$, $z = 3.75$, $p < 0.001$).

The duration analysis on tokens with positive VOT showed a significant effect of stimulus category, such that productions in response to [k] stimuli had longer VOT than those in response to [g] ($\beta = 54.57$, $SE = 9.59$, $\chi^2(1) = 15.55$, $p < 0.001$). No other effects were significant.

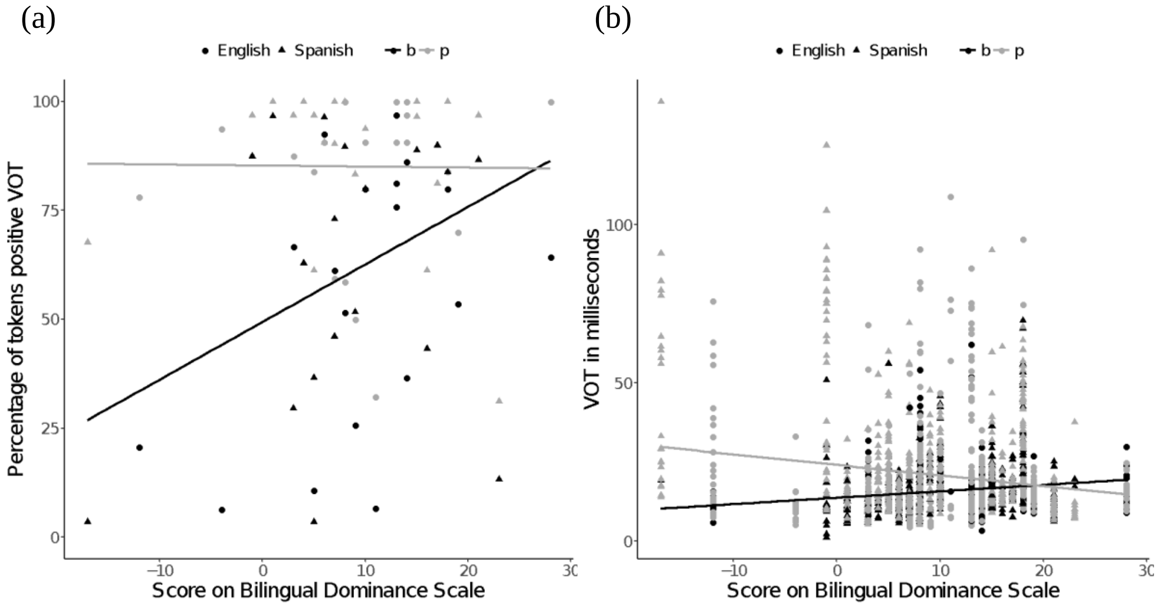

Figure 4. In (a) percentage of bilinguals' tokens produced with positive VOT in response to negative and short-lag VOT stimuli ([b] and [p] respectively) as a function of BDS; in (b) duration of positive VOT productions in response to the same set of stimuli; language mode is also noted.

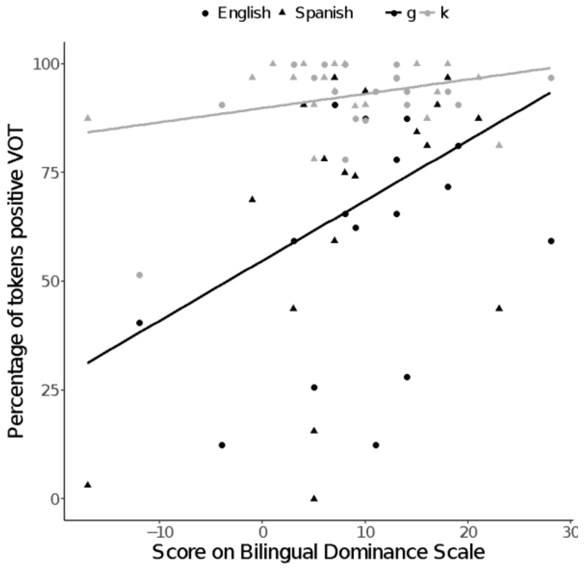

Figure 5. Percentage of bilinguals' tokens produced with positive VOT in response to negative and short-lag VOT stimuli ([g] and [k] respectively) as a function of BDS; language mode is also noted.

2.2.2. Bilinguals vs. monolinguals

Figure 6 presents the results for bilinguals and monolinguals together; the data for the bilinguals is the same as in Figure 3 but collapsed across language mode.

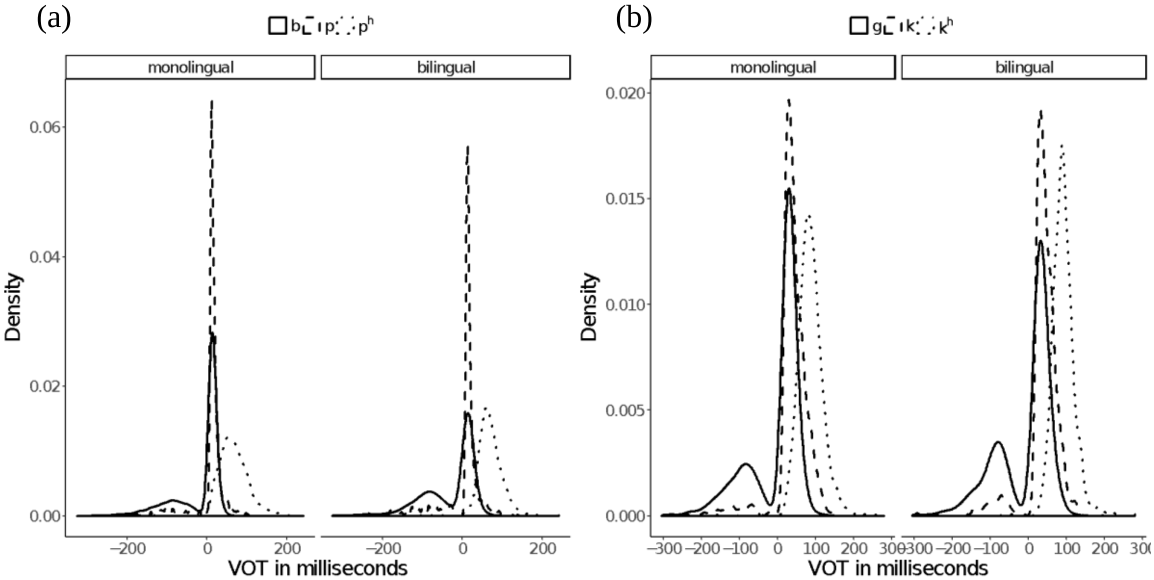

Figure 6. Distributions of monolinguals' and bilinguals' VOT values in the production experiment by language background and stimulus category for (a) bilabial stops and (b) velar stops.

The analysis of bilabials for negative vs. positive VOT showed a significant effect of stimulus category, such that responses to [p] stimuli were more likely to have positive VOT than responses to [b] stimuli (β = 1.81, *SE* = 0.16, $\chi^2(1)$ = 73.90, $p < 0.001$). There was also a trending effect of language background (β = -0.69, *SE* = 0.39, $\chi^2(1)$ = 3.07, $p = 0.08$). For the duration analysis on positive tokens, there was a significant effect of stimulus category, such that the VOT of tokens produced in response to [p] stimuli was longer than that produced in response to [b] stimuli (β = 4.85, *SE* = 0.99, $\chi^2(1)$ = 10.26, $p = 0.001$). No other effects were significant.

The analysis of velars for negative vs. positive VOT showed a significant effect of stimulus category, such that [k] stimuli were more likely to be responded to with positive VOT than [g] stimuli (β = 2.30, *SE* = 0.19, $\chi^2(1)$ = 82.13, $p < 0.001$). No other effects were significant.

The duration analysis on tokens with positive VOT showed a significant effect of stimulus category, such that VOT productions in response to [k] stimuli had longer durations than those produced in response to [g] stimuli (β = 11.99, *SE* = 2.18, $\chi^2(1)$ = 9.24, $p = 0.002$). No other effects were significant.[2]

2.3. Interim discussion

The hypothesis for the production task was that the stimuli would be either imitated, and thus reproduced faithfully by bilinguals (who have at their disposal the whole gamut of VOT

[2] Analyses comparing the production of monolinguals and bilinguals in response to short-lag vs. long-lag stimuli showed that VOT productions were longer in response to the long-lag than the short-lag stimuli (bilabial: β = 48.82, *SE* = 3.06, $\chi^2(1)$ = 19.15, $p < 0.001$; velar: β = 41.87, *SE* = 2.86, $\chi^2(1)$ = 17.30, $p < 0.001$). No other effects were significant for either model.

values), or categorized phonologically, in which instance either the categorization of English or that of Spanish would prevail, based on language dominance or mode. The results suggest that all participants treated the task as imitation: even the monolingual participants produced prevoiced stops in response to stimuli with negative VOT, something that is rare in English (Davidson, 2016). Beach et al. (2001) found similar results with monolingual English speakers and Greek-English bilinguals using an imitation task: both groups responded to prevoiced Thai stops with prevoiced productions, though neither group used prevoicing when reading English.[3] In addition, the bilinguals in the present study showed a dominance effect: in response to stimuli with negative VOT, more English-dominant bilinguals produced tokens typical of English (i.e., with short positive VOT), while more Spanish-dominant participants produced shorter VOT (for bilabials) and fewer tokens with positive VOT. These dominance effects are comparable to those reported in MacKay et al. (2001) and Beach et al. (2001) for Italian-English and Greek-English bilinguals respectively. On the other hand, our bilingual participants were not influenced by language mode, which has been considered critical in bilingual production and perception (Elman et al., 1977; Grosjean, 2001). We return to this point in section 6.

The comparisons between monolingual and bilingual speakers showed no differences *in the aggregate*. Closer inspection, however, indicates that there were *within* group differences between monolinguals and bilinguals. As illustrated in Figure 7, which shows individual responses to stimuli with prevoicing (i.e., [b] and [g]), there is large variation in the productions for both groups, in line with previous findings (Theodore et al., 2009, on English monolinguals; Beach et al., 2001, on bilinguals). What is of interest, however, is the difference in patterns between monolingual and bilingual speakers; 11 bilinguals showed unimodal distributions with negative VOT values (e.g., BS_05, BS_19) or bimodal distributions with predominantly negative values (e.g., BE_33, BE_35); neither pattern is present in the monolingual data. In addition, fewer bilinguals than monolinguals failed to produce prevoiced tokens. Examining the data at the level of the individual indicates that although the observed variability can lead to non-significant statistical differences between groups, speakers in each group adopt different production strategies, with bilinguals showing effects from their two languages, as well as greater inter-speaker variability than the monolingual group.

3. Experiment 2: Perception - ABX

3.1. Method

3.1.1. Participants

Twenty monolinguals and twenty bilinguals from the 80 participants in the production study took part in the ABX task. The bilinguals had an average BDS score of 11.0 (*SD* = 5.87).

[3] Greek has prevoiced and voiceless unaspirated stops, while Thai has prevoiced, voiceless unaspirated and voiceless aspirated stops. Thus, the experiment of Beach et al. (2001) is analogous to our own.

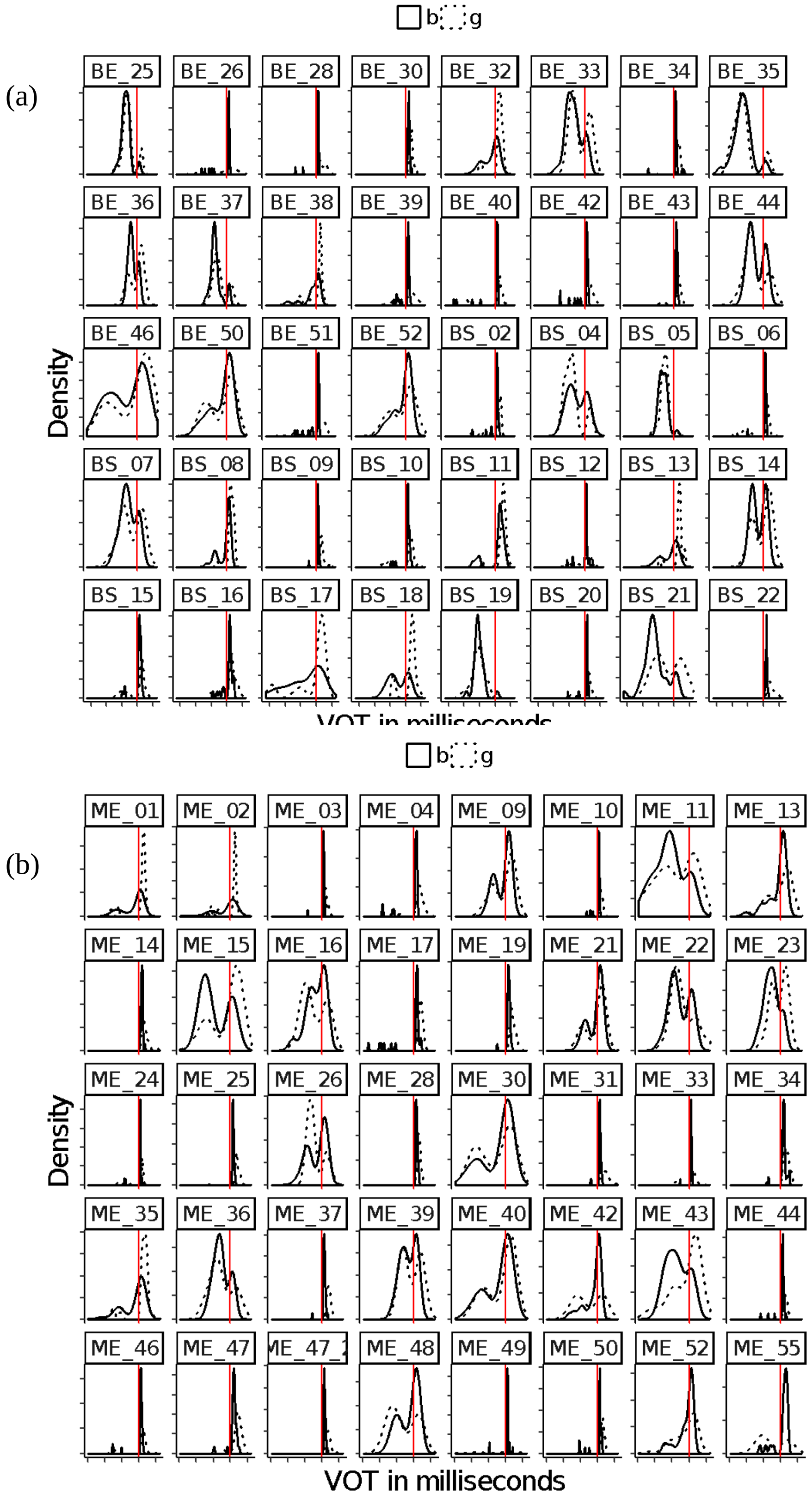

Figure 7. Productions by individual speakers for (a) bilinguals and (b) monolinguals for the two negative VOT stimulus categories, [b] and [g]. A line at 0 ms is marked to compare positive vs. negative VOT durations. Note that the y-axis scale is not the same for all speakers.

3.1.2. Stimuli

The same set of words was used as in Experiment 1. The experiment included 288 trials each consisting of a series of three stimuli. The third stimulus was always the same word as either the first or second stimulus, with matches counterbalanced across trials. The first two stimuli were produced by one speaker and the third by the other, with speaker order counterbalanced across trials. Of the 288 trials, 48 were test trials in which the first two stimuli differed only in VOT duration (24 bilabial, 24 velar). These trials were constructed such that participants heard each test contrast (e.g., [b] vs. [p], [b] vs. [p^h], etc.) eight times. The VOT category of the first stimulus in each trial and whether it was the correct answer were also counterbalanced. The other 240 trials were fillers in which the stimuli differed in more than VOT.

3.1.3. Procedure

Experiment 2 was administered immediately following Experiment 1, which thus served as a form of familiarization with the stimuli. The stimuli were presented over headphones using SuperLab Pro 4.5 (Cedrus Corporation, 2011). Participants were told they would hear words from the same language they had heard in the first experiment. They were asked to focus on a fixation cross during which time they would hear three words; the ISI between the appearance of the cross and the first stimulus, and between successive stimuli in a trial was 500 ms. The fixation cross went away as soon as the last stimulus finished playing; participants then had to press a key to indicate whether the third word was the same as the first ("z" key) or the second word ("m" key). A reminder of which key to use for each response was presented as soon as the third stimulus finished playing. Monolinguals were tested in English. Bilinguals were tested either in English or Spanish (*n* = 10 per mode). The mean BDS score for the English-mode group was 12.20 (*SD* = 4.42), and for the Spanish-mode group 9.80 (*SD* = 7.07) (*t* = 0.91, *p* = 0.38).

3.1.4. Analyses

Two analyses, one on bilinguals to test for effects of language mode and dominance, and one comparing monolinguals to bilinguals, were run on negative vs. short-lag and short-lag vs. long-lag VOT. For the analysis of bilinguals, a generalized linear mixed effects model was run with accuracy (correct, incorrect) as the dependent variable. Contrast (negative vs. short-lag VOT, short-lag vs. long-lag VOT), language mode, and BDS score were included as fixed effects together with two interactions, contrast × language mode and contrast × BDS score. Contrast and language mode were contrast coded; BDS score was coded as a numeric variable. Participant was included as a random intercept and a random slope by contrast uncorrelated with the random intercept; item was included as a random intercept.

For the analysis of bilinguals vs. monolinguals, similar models were run, the main difference being that they included contrast and language background (monolingual, bilinguals) and their interactions as fixed effects. Participant was included as a random intercept and a random slope by contrast; item was included as a random intercept and a random slope by language background uncorrelated with the random intercept.

For all models, these were the maximal random effects structures that would converge (Barr et al., 2013). Significance testing was done with model comparison with alpha set to 0.05.

3.2. Results

3.2.1. Bilinguals

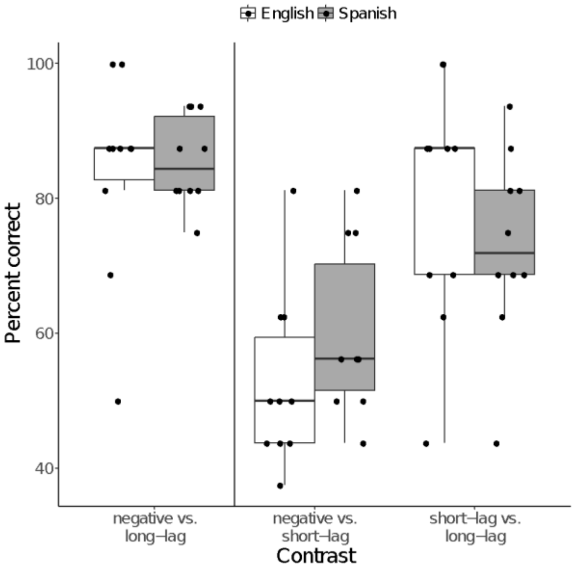

Figure 8. Bilinguals’ performance on the ABX task by contrast and language mode.

For bilinguals, the only significant effect was the interaction of contrast and BDS score (β = 0.10, *SE* = 0.04, $\chi^2(1)$ = 7.20, *p* = 0.007); see Figure 8. For negative vs. short-lag VOT, a higher BDS score resulted in lower accuracy (β = -0.05, *SE* = 0.02, *z* = -2.60, *p* = 0.009); for short-lag vs. long-lag, higher BDS score resulting in *higher* accuracy (β = 0.05, *SE* = 0.02, *z* = 2.28, *p* = 0.02); see Figure 9.

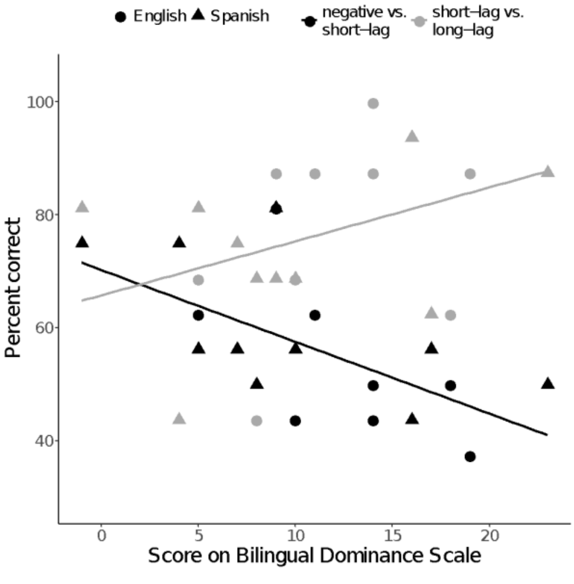

Figure 9. Bilinguals’ performance on the ABX task by contrast and score on the Bilingual Dominance Scale; language mode is also noted.

3.2.2. Bilinguals vs. monolinguals

Figure 10 presents the results for bilinguals and monolinguals together; the data for the bilinguals is the same as in Figure 9 but collapsed across language mode. There was a significant effect of contrast such that participants were more accurate on short-lag vs. long-lag VOT than negative vs. short-lag VOT ($\beta = 0.77$, $SE = 0.21$, $\chi^2(1) = 11.81$, $p < 0.001$). No other effects were significant.

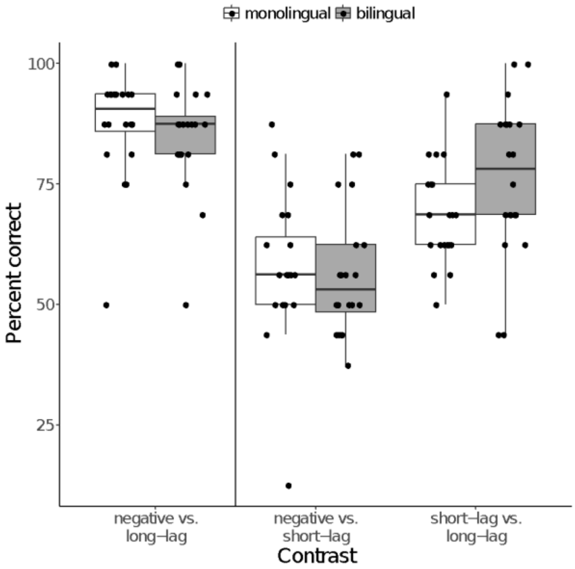

Figure 10. Monolinguals' and bilinguals' performance on the ABX task by contrast.

3.3. Interim discussion

Our hypothesis was that the phonological categorization required in ABX would present a conflict for bilinguals with respect to stimuli with short-lag VOT. This hypothesis was supported by their significantly higher accuracy in negative vs. long-lag and short-lag vs. long-lag VOT, relative to their low accuracy for negative vs. short-lag VOT, the contrast involving the two VOT types that belong to the same phonological category in English.

Although the results suggest that bilinguals did not differ from monolinguals, the interaction between accuracy and BDS also indicates an effect of language dominance similar to that in the production task. More English-dominant participants were less accurate in the critical negative- vs. short-lag VOT contrast (the contrast found only in Spanish), relative to those who were more balanced. In contrast, the English-dominant participants were more accurate at short- vs. long-lag VOT, categories relevant for English but irrelevant for Spanish. The interaction in Figure 9 indicates that the more balanced bilinguals were equally good at both contrasts, but with average performance in both.

4. Experiment 3: Perception - AX

4.1. Method

4.1.1. Participants

The participants were the other half of those who took part in the production task: 20 bilinguals, and 20 monolinguals. The bilinguals had an average BDS score of 7.45 ($SD = 10.57$), which was not statistically different from that in Experiment 2 ($t = 1.31$, $p = 0.20$).

4.1.2. Stimuli

The same set of words was used as in Experiments 1 and 2. The experiment included 288 trials consisting of two stimuli produced by different speakers, with speaker order counterbalanced across trials. For half of the trials the two stimuli were the same word. In the "same" set, half the trials included test items and the other half fillers (see section 2.1.2). Listeners heard each "same" pair in two trials (for which different tokens were used). Of the 144 "different" trials, 36 were test items differing only in the VOT of the initial stop; for these. all possible pairings were used (e.g., [bɛma]-[pɛma], [bɛma]-[pʰɛma], [pɛma]-[pʰɛma]), with the order of stimuli counterbalanced across trials. The rest of the "different" trials were fillers constructed so that each word was heard an equal number of times through the course of the experiment.

4.1.3. Procedure

The experiment procedure was the same as in Experiment 2, except that listeners heard two words per trial. Listeners were told that the words would not sound exactly the same as they were produced by two different speakers. As in Experiment 2, bilinguals were tested either in English or Spanish mode (*n* = 10 per mode). The mean BDS score for the English-mode group was 7.50 (*SD* = 10.78), and for the Spanish-mode group 7.40 (*SD* = 10.94) (*t* = 0.02, *p* = 0.98).

4.2. Results

4.2.1. Bilinguals

Figure 11 shows the results for bilinguals on same and different trials. Same trials, in which performance was at ceiling, were not analyzed further. The analysis for the "different" trials, which was the same as in Experiment 2, showed only a trending interaction of contrast and language mode (β = -1.20, *SE* = 0.62, $\chi^2(1)$ = 3.36, *p* = 0.07). No other effects were significant.

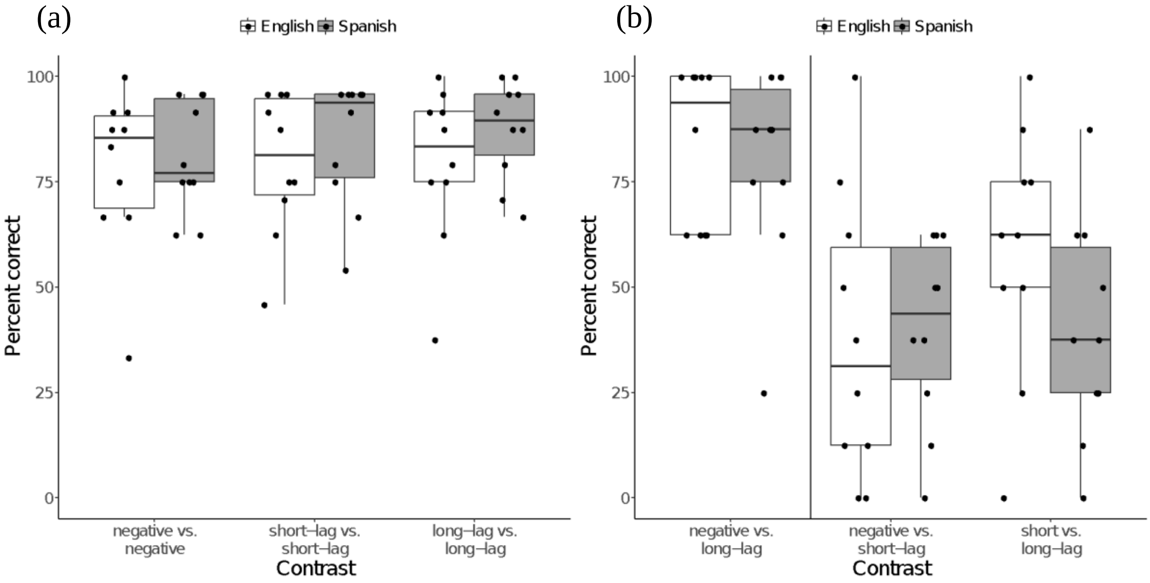

Figure 11. Bilinguals' performance on the AX task by contrast and language mode for (a) same trials and (b) different trials.

4.2.2. Bilinguals vs. monolinguals

Figure 12 shows the results for bilinguals vs. monolinguals on same and different trials. Neither of the main effects nor the interaction were significant.

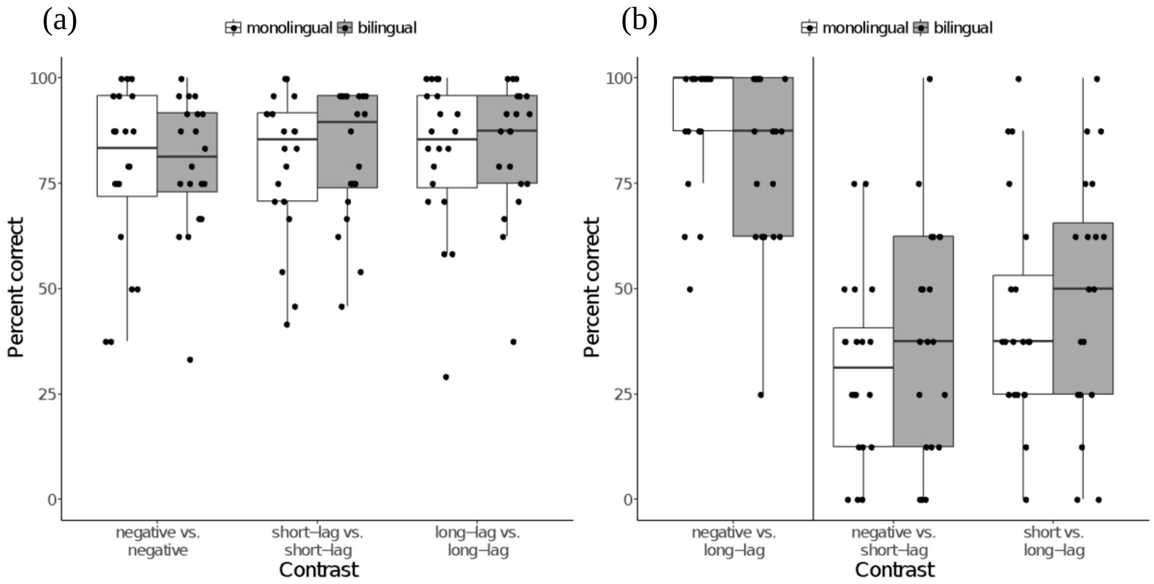

Figure 12. Monolinguals' and bilinguals' performance on the AX task by contrast and language background for (a) same trials and (b) different trials.

4.3. Interim discussion

We had hypothesized that the AX task, similarly to Experiment 1, would be relatively simple for bilinguals since it requires sensitivity to phonetic detail available to them from English and Spanish. Accuracy was very low, however, especially for the "different" trials. This indicates a strong conservative bias, possibly due to the fact that participants were cautioned the two stimuli in each trial would not be identical; this may have made them treat audible differences as speaker-related, responding with "different" only to trials involving maximally distinct stimuli, those that juxtaposed negative with long-lag VOT.

5. Production vs. perception

A final analysis was conducted to compare production and perception.

5.1. Analyses

For each participant, the median VOT duration of their tokens produced in Experiment 1 in response to each word was calculated. Median was a better approximation of "average" than mean as distributions were not always normal; for VOTs in response to negative stimuli the median may have been pulled from a bimodal distribution. The difference in median VOT was computed for all stimuli pairs in the perception experiment to which the participant was assigned (AX for the AX task; AB for the ABX task). For example, if a trial in the participant's perception task compared [bɛm] and [pɛm], and in the production task s/he had a median VOT duration of -82.13 ms in response to [bɛm] and a median VOT duration of 10.35 ms in response to [pɛm], the difference between them was 92.48 ms (-82.13 − 10.35). This was done for each of the word pairings in the perception tasks for the two key contrasts (negative vs. short-lag VOT, short-lag vs. long-lag VOT). We expected that participants showing larger differences in production would be more accurate in perception (cf. Beach et al., 2001).

To test this hypothesis, four generalized linear mixed effects models were run, one for monolinguals and one for bilinguals in the ABX task and AX task respectively. Monolinguals and bilinguals were tested separately due to model convergence issues. The dependent variable was accuracy (correct, incorrect). The fixed effects were contrast (negative vs. short-lag VOT, short-lag vs. long-lag VOT) and VOT duration difference between words. Contrast was coded with contrast coding and VOT duration difference was included as a numeric variable. Participant and item were included as random intercepts. This was the maximal random effects structure that would converge (Barr et al., 2013). Significance testing was done with model comparison with alpha set to 0.05.

5.2. Results

5.2.1. Production and ABX

As the effect of contrast has already been reported in section 3.2, only effects regarding the duration differences are reported here. For the bilinguals, no effects regarding duration difference were found (see Figure 13). For monolinguals, however, a larger difference in production resulted in higher accuracy in perception (β = 0.007, *SE* = 0.003, $\chi^2(1)$ = 6.73, *p* = 0.01). There was also a significant interaction of contrast and duration difference (β = 0.01, *SE* = 0.006, $\chi^2(1)$ = 4.07, *p* = 0.04). For negative vs. short-lag VOT, there was no effect of duration difference, but for short-lag vs. long-lag VOT a larger difference in production correlated with higher accuracy in perception (β = 0.01, *SE* = 0.005, *z* = 3.24, *p* = 0.001).

5.2.2. Production and AX

For monolinguals, the same result was found as for the ABX task (see Figure 14): larger duration differences in production resulted in higher accuracy in AX (β = 0.01, *SE* = 0.006, $\chi^2(1)$ = 6.50, *p* = 0.01). Investigation of the interaction of contrast and duration difference showed no effect for negative vs. short-lag VOT (β = 0.03, *SE* = 0.01, $\chi^2(1)$ = 7.71, *p* = 0.005), but for short-lag vs. long-lag VOT, a larger difference in production resulted in higher accuracy in perception (β = 0.02, *SE* = 0.007, *z* = 3.21, *p* = 0.001).

For bilinguals, there was a significant interaction of contrast and duration difference (β = 0.02, *SE* = 0.01, $\chi^2(1)$ = 5.18, *p* = 0.02). For short-lag vs. long-lag VOT, a larger difference resulted in higher accuracy (β = 0.02, *SE* = 0.007, *z* = 3.17, *p* = 0.002), but for negative vs. short-lag VOT, a larger difference resulted in *lower* accuracy (β = -0.008, *SE* = 0.003, *z* = -2.12, *p* = 0.03).

5.3. Interim discussion

The comparison of production and perception uncovered differences between monolinguals and bilinguals. For monolinguals, production and perception were highly connected: those who produced VOT categories more distinctly could detect these differences in perception, whether processing at the phonetic (AX) or phonological level (ABX). This is unsurprising, since for monolinguals phonetic categories are mapped to phonological ones in a straightforward manner.

For bilinguals, production was not as well associated with perception. For ABX participants, production and perception did not correlate, indicating that although the three VOT categories are available to them at the phonetic level, resulting in accurate production, the conflicting phonological classification of short-lag VOT and the irrelevance of long-lag VOT for Spanish did not allow them to take advantage of this familiarity. For AX participants, results depended on contrast. For short-lag vs. long-lag, the relation was straightforward: greater contrast in production correlated with greater sensitivity in perception (as with monolinguals). The puzzling negative correlation between production and accuracy in negative vs. short-lag VOT may be evidence of the conservative bias discussed in section 4.3 (note that the effect is present in the monolingual data too, though without reaching statistical significance; see Figure 14).

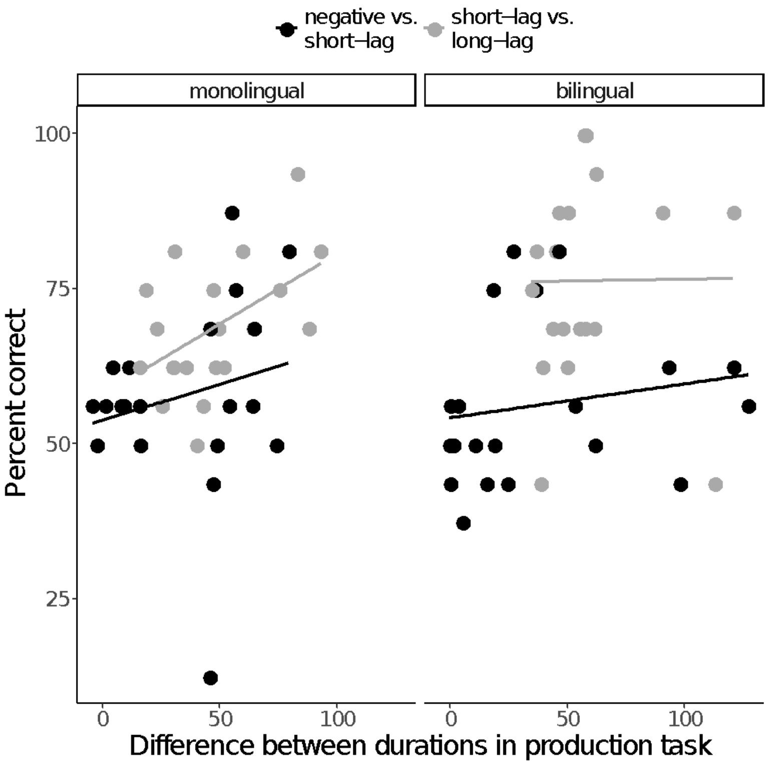

Figure 13. Correlation between production and perception for ABX task for monolinguals and bilinguals.

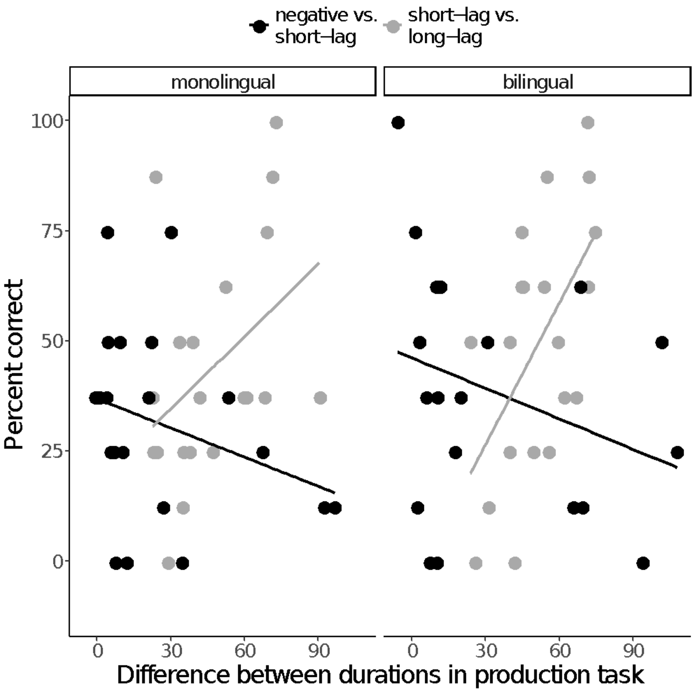

Figure 14. Correlation between production and perception for AX task for monolinguals and bilinguals.

6. General discussion

We investigated the production and perception of stops with negative, short-lag and long-lag VOT among English monolinguals and Spanish-English bilinguals, to shed light on the connection between production and perception and assess the role of language mode and language dominance in both. We focused on stops with short-lag VOT because they are phonologically classified in conflicting ways in Spanish and English, and we wished to probe the issue of phonetic vs. phonological processing among bilinguals.

Our bilingual group indicated a language dominance effect. In production, more balanced bilinguals better differentiated negative from short-lag VOT, producing more instances of the former in response to [b] and [g] as compared to [p] and [k], and having a larger duration difference between short-lag VOT tokens produced in response to [b] than [p]. Dominance also mattered in the ABX task: balanced bilinguals performed similarly in negative vs. short-lag, and short- vs. long-lag VOT, and were not very accurate in either; English-dominant bilinguals, on the other hand, did worse on the former than the latter contrast, as would be expected if their processing was based on English.

Using BDS allowed us to incorporate dominance in our statistical modelling and show that dominance is gradient and can have gradient effects on bilingual production and perception. This means that matching bilinguals on the basis of characteristics such as age of acquisition (MacKay et al., 2001; Dunn & Fox Tree, 2009), does not necessarily result in homogeneous group performance. Thus, considering group results may miss the granularity with which factors like dominance affect bilingual production and perception.

Additionally, our production results indicate inter-speaker variation (see also section 2.3). Although *intra*-speaker variability in production is to be expected for VOT (Theodore et al., 2009), the inter-speaker variation of the bilingual group goes beyond the mundane: some participants were very close to English monolinguals, others to Spanish monolinguals, and still others vacillated between the two, presenting strongly bimodal distributions of negative and short-lag VOT in response to negative VOT stimuli. This inter-speaker variation pertained to the realization of categories that result in conflicting phonological classifications across the two languages. Such categories may be critical for understanding bilingualism: the learning of new phonetic categories and the restructuring of existing ones in response to exposure to a second language is well documented (Best & Strange, 1992; Flege, 1995; MacKay et al., 2001; Fowler et al., 2008). However, conflicting phonological categorizations like those discussed here clearly result in different strategies for individual speakers even when their linguistic experiences are comparable by a number of metrics. In short, we see different patterns of responses adopted within the same group of bilinguals, *as well as* gradience within each pattern. These different strategies deserve greater attention, as they may well be the reason why studies on bilingualism do not provide a consistent picture. In our view, they can best be explained in terms of activation and suppression of linguistic subsystems (Green, 1998): some bilinguals can suppress one language effectively and consistently, while others cannot. This is a plausible explanation for the behavior of the more balanced bilinguals, the group more likely to have difficulties suppressing one language: these participants may have operated sometimes in English and sometimes in Spanish mode, particularly when faced with conflicting categorization options. Such switches could apply over and above gradient effects which may relate to the restructuring of old and the development of new phonetic categories in response to second language input. This dual model would allow for both categorical and gradient effects in production and perception.

An additional possibility is that individuals simply have different skills when it comes to attending to phonetic detail. Such differences are present among monolinguals and bilinguals alike (cf. Beach et al., 2001; Lengeris & Hazan, 2011; Yu, Abrego-Collier, & Sonderegger, 2013), and most likely operate over and above any effects of bilingualism *per se* (Beach et al., 2001). Such individual variation may also explain why the bilingual participants, as a group, did not differ statistically from monolinguals.

Differences between groups were evident in the connection between production and perception. For monolinguals, production and perception correlated at least for categories pertinent to English phonology, namely short- vs. long-lag VOT. For bilinguals, production and perception were more distinct, as also shown by Beach et al. (2001), reflecting the conflict created when phonetic categories are classified differently in the phonology of the speaker's two languages. Together, the results of the monolingual and bilingual participants indicate that exposure to and familiarity with phonetic categories may not be as critical as phonological classification, especially in perception. Furthermore, the results reflect the importance of using a battery of tasks, and testing both production and perception before conclusions are drawn, as clearly the two are not always connected in a straightforward manner.

Unlike dominance, our experiments did not show effects of language mode. Although in the perception tasks this could be partly attributed to low power, this does not apply in production. Though these results seemingly question the importance of language mode, it is possible they relate to the nature of the tasks involved. First, task instructions explicitly mentioned that the stimuli were from a language unknown to the participants; thus language mode may have been ignored, with participants relying on their dominant language (cf. Beach et al., 2011). Second, the task was completed in an English-dominant setting (a University campus in the USA), so the mode manipulation may have not been sufficient to push participants out of English mode. Finally, dominance may have overshadowed mode. Language dominance is recognized as a critical component of bilingual production and perception (e.g., Flege & Eefting, 1987; Flege et al., 2002; Olson, 2013), but its combination with mode has not been investigated to a large extent. Our results indicate that their interaction deserves more attention.

Finally, our results suggest that the distinction between phonetic and phonological processing is valid and should be taken into consideration: while bilinguals have at their disposal the phonetic categories of their two languages, consistent with some degree of constant activation of both, for tasks that require phonological abstraction and categorization, they need to operate in one language at a time. In turn, this means that for phonological processing, one language must be actively suppressed. This language may not remain the same throughout a task, and could be determined by a number of factors, including dominance and mode (cf. Mack, 1989; Beach et al., 2001).

7. Conclusion

Bilingual speech production and perception are more complex than often suggested and driven by a number of factors, including language dominance, and individual variation in dealing with conflicts arising from features of a bilingual's languages. The differences found between monolinguals and bilinguals in the connection between production and perception further suggests that it is important to consider both in future research. In short, language dominance, individual variation, and the relationship between production and perception deserve significantly more attention in the study of bilingualism.

References

Abramson, A. S., & Whalen, D. H. (2017). Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions. *Journal of Phonetics*, *63*, 75–86.

Balukas, C., & Koops, C. (2014). Spanish-English bilingual voice onset time in spontaneous code-switching. *International Journal of Bilingualism*, *19*(4), 423–443.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Beach, E. F., Burnham, D., & Kitamura, C. (2001). Bilingualism and the relationship between perception and production: Greek/English bilinguals and Thai bilabial stops. *The International Journal of Bilingualism*, *5*(2), 221-235.

Beddor, P. (2017). Speech Perception in Phonetics. *Oxford Research Encyclopedia of Linguistics.* Retrieved 2 Nov. 2017, from http://linguistics.oxfordre.com/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-62.

Best, C. T., & Strange, W. (1992). Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of Phonetics*, *20*(3), 305-330.

Bills, G. D., Chávez, E. H., & Hudson, A. (1995). The geography of language shift: Distance from the Mexican border and Spanish language claiming in the Southwestern US. *International Journal of the Sociology of Language*, *114*(1), 9-28.

Cedrus Corporation. (2011). *SuperLab 4.5*. San Pedro, CA.

Cho, T. & Ladefoged, P. (1999). Variation and universals in VOT: Evidence from 18 languages. *Journal of Phonetics*, *27*(2), 207-229.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1992). The monolingual nature of speech by bilinguals. *Cognitive Psychology, 24*(3), 381–410.

Davidson, L. (2016). Variability in the implementation of voicing in American English obstruents. *Journal of Phonetics, 54*, 35–50.

Dmitrieva, O., Llanos, F., Shultz, A. A., & Francis, A. L. (2015). Phonological status, not voice onset time, determines the acoustic realization of onset f0 as a secondary voicing cue in Spanish and English. *Journal of Phonetics, 49*, 77-95.

Docherty, G. J. (1992). *The Timing of Voicing in British English Obstruents*. Berlin/New York: Foris.

Dum-Tragut, J. (2009). *Armenian: Modern Eastern Armenian* (Vol. 14). John Benjamins Publishing.

Dunn, A. L., & Fox Tree, J. E. (2009). A quick, gradient Bilingual Dominance Scale. *Bilingualism: Language and Cognition*, *12*(3), 273–289.

Elman, J. L., Diehl, R. L., & Buchwald, S. E. (1977). Perceptual switching in bilinguals. The *Journal of the acoustical Society of America*, *62*(4), 971-974.

Flege, J. E. (1995). Second-language speech learning: Theory, findings and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–273). Timonium, MD: York Press.

Flege, J. E., & Eefting, W. (1987). Production and perception of English stops by native Spanish speakers. *Journal of Phonetics*, *15*, 67-83.

Flege, J. E., & Eefting, W. (1988). Imitation of a VOT continuum by native speakers of English and Spanish: evidence for phonetic category formation. *Journal of the Acoustical Society of America*, *83*(2), 729–40.

Flege, J. E., MacKay, I. R. A., & Piske, T. (2002). Assessing bilingual dominance. *Applied Psycholinguistics, 23*, 567–598.

Fowler, C. A., Sramko, V., Ostry, D. J., Rowland, S. A., & Hallé, P. (2008). Cross language phonetic influences on the speech of French–English bilinguals. *Journal of Phonetics, 36*(4), 649-663.

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, *1*(2), 67–81.

Grijalva, C., Piccinini, P., & Arvaniti, A. (2013). The vowel spaces of Southern California English and Mexican Spanish as produced by monolinguals and bilinguals. *POMA* (*Proceedings of Meetings on Acoustics*), *19*(1), 060088.

Grosjean, F. (1989). Neurolinguists beware! The bilingual is not two monolinguals in one person. *Brain and Language, 36*, 3-15.

Grosjean, F. (1998). Studying bilinguals: Methodological and conceptual issues. *Bilingualism: Language and Cognition, 1*(2), 131-149.

Grosjean, F. (2001). The bilingual's language modes. In Nicol, J. (Ed.). *One Mind, Two Languages: Bilingual Language Processing* (pp. 1-22). Oxford: Blackwell. Also in Li Wei (Ed.). *The Bilingual Reader* (2nd edition). London: Routledge, 2007.

Keating, P. A. (1984). Phonetic and phonological representation of stop consonant voicing. *Language, 60*(2), 286-319.

Kessinger, R. H., & Blumstein, S. E. (1997). Effects of speaking rate on voice-onset time in Thai, French, and English. *Journal of Phonetics*, *25*(2), 143–168.

Lengeris, A., & Hazan, V. (2011). The effect of native vowel processing ability and frequency discrimination acuity on the phonetic training of English vowels for native speakers of Greek. *Journal of the Acoustical Society of America, 128*, 3757-3768.

Lipski, J. M. (2008). *Varieties of Spanish in the United States*. Georgetown University Press.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384–422.

Mack, M. (1989). Consonant and vowel perception and production: Early English-French bilinguals and English monolinguals. *Perception & Psychophysics, 46*(2), 187-200.

MacKay, I. R., Flege, J. E., Piske, T., & Schirru, C. (2001). Category restructuring during second-language speech acquisition. *Journal of the Acoustical Society of America, 110*(1), 516-528.

Magloire, J., & Green, K. P. (1999). A Cross-Language Comparison of Speaking Rate Effects on the Production of Voice Onset Time in English and Spanish. *Phonetica*, *56*(3-4), 158–185.

McGuire, G. (2010). A Brief Primer on Experimental Designs for Speech Perception Research. Available at https://people.ucsc.edu/~gmcguir1/experiment_designs.pdf.

Nakai, S., & Scobbie, J. M. (2016). The VOT category boundary in word-initial stops: Counter-Evidence against rate normalization in English spontaneous speech. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *7*(1), 13.

Olson, D. J. (2013). Bilingual language switching and selection at the phonetic level: Asymmetrical transfer in VOT production. *Journal of Phonetics, 41*(6), 407-420.

Piccinini, P., & Arvaniti, A. (2015). Voice onset time in Spanish–English spontaneous code-switching. *Journal of Phonetics, 52*, 121-137.

Pisoni, D. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception and Psychophysics, 13*(2), 253-260.

Raphael, L. J., Borden, G. J., & Harris, K. S. (2007). *Speech Science Primer: Physiology, Acoustics, and Perception of Speech*. Lippincott Williams & Wilkins.

Stuart-Smith, J., Sonderegger, M., Rathcke, T., & Macdonald, R. (2015). The private life of stops: VOT in a real-time corpus of spontaneous Glaswegian. *Laboratory Phonology*, *6*(3-4), 505-549.

Sundara, M., Polka, L., & Baum, S. (2006). Production of coronal stops by simultaneous bilingual adults. *Bilingualism: Language and Cognition, 9*(1), 97-114.

Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice-onset-time: Contextual influences. *The Journal of the Acoustical Society of America, 125*(6), 3974-3982.

Williams, L. (1977). The perception of stop consonant voicing by Spanish-English bilinguals. *Perception & Psychophysics, 21*(4), 289-297.

Yu, A., Abrego-Collier, C., & Sonderegger, M. (2013). Phonetic imitation from an individual-difference perspective: Subjective attitude, personality and "autistic" traits. *PLOS ONE* 8(9): e74746. Published: September 30, 2013.