Phonetic category formation is perceptually driven during the early stages of adult L2 development

Joseph V. Casillas[1]

[1] Rutgers University

Author Note

I would like to thank Miquel Simonet, Yukari Hirata, Joseph Kern, and three anonymous reviewers for their candid comments and suggestions on an earlier version of this manuscript. All errors are mine alone. This work was partially funded by the Comanche Nation Higher Education Grant (Nʉmʉnʉʉ), as well as the University of Arizona Graduate and Professional Student Council Research and Project Grant (RSRCH-702FY15).

Correspondence concerning this article should be addressed to Joseph V. Casillas, Rutgers University - Department of Spanish and Portuguese, 15 Seminary place, New Brunswick, New Jersey, 08904, USA . E-mail: joseph.casillas@rutgers.edu

Abstract

Research on the acquisition of L2 phonology in sequential language learners has stressed the importance of language use and input as a means to accurate production and perception; however, the two constructs are difficult to evaluate and control. This study focuses on the role of language use during the initial stages of development of phonetic categories related to stop voicing and analyzes the relationship between production and perception. Native English speaking late learners of Spanish provided production/perception data on a weekly basis throughout the course of a 7-week immersion program in which L1 use was prohibited. The production/perception data were analyzed using generalized linear mixed effects models. Generalized additive mixed models were used to analyze and compare the learning trajectories of each modality. The analyses revealed phonetic learning in both production and perception over the course of the program. Perception gains paralleled those of native bilinguals by the conclusion of the program and preceded production gains. This study is novel in that it provides production/perception data in a semi-longitudinal design. Moreover, the beginning adult learners are examined in a learning context in which L1 use was minimal and L2 input was maximized. Taken together, the experiments suggest that L2 phonetic category formation can occur abruptly, at an early stage of development, is perceptually driven, and appears to be particularly fragile during the initial stages of learning.

Keywords: Sequential language learning, SLA, Production, Perception, Stop voicing

Word count: 12644

Phonetic category formation is perceptually driven during the early stages of adult L2 development

Introduction

A common finding in the second language (L2) speech literature is that adults who learn another language typically retain a non-native accent (Caramazza et al., 1973; Flege, 1981, 1987a; Fowler et al., 2008; Oyama, 1976; Pallier et al., 1997; Sundara & Polka, 2008, among others). The phonetic consequences of sequential language learning—acquiring an L2 after, rather than at the same time as, the L1—are traditionally associated with speech production. Some researchers, however, refer to L2 learners as perceiving speech with an accent as well (Escudero, 2005; Strange, 1995). That is to say, L2 speech perception can also differ from native listening. There is a dearth of research regarding the relationship between production and perception in adult L2 learning and how the two modalities are affected by L2 input and L2 use.

The present work is concerned with understanding how late learners manage to acquire L2 sound categories and the nature of their development in reference to input and L1/L2 use. Additionally, this work explores the interface between speech perception and speech production in beginning adult learners. An L2 can be learned formally (i.e., in an L2 classroom), informally (i.e., in a naturalistic context), or both formally and informally in an immersion type context (Saville-Troike, 2005). The present work focuses on the L2 learning that takes place in the latter. Data were collected from learners in a domestic immersion context—i.e., foreign language immersion in their country of origin, the U.S.—in which they were required to minimize the use of their L1 and received large amounts of L2 input. Specifically, this work examines the initial stages of L2 production and perception in a group of adult late learners that took part in a Spanish domestic immersion program in which L1 use was prohibited.

Background

Input and use in L2 learning

While age effects have duly garnered the attention of second language acquisition (SLA) researchers, we know now that early exposure alone cannot explain L2 outcomes in all cases (See Pallier et al., 1997; Sebastián-Gallés & Soto-Faraco, 1999; Sebastián-Gallés et al., 2005; Sebastián-Gallés, 2006, among others). A possible explanation may lie in the nature of the input learners receive. For instance, in an investigation of the acquisition of English /d/-/ð/, Sundara, Polka, & Genesee (2006) determined that variable realizations of English /ð/ in the input of young French/English bilinguals may have delayed their acquisition of a "functional" /d/-/ð/ contrast (p. 382). This suggests that the difficulties in perception/production of some learners might be a reflection of the input to which they are exposed.

The role of input in the production/perception of adult learners has also been studied, albeit to a lesser extent. For instance, Flege & Liu (2001) investigated length of residency (LOR) and input in a group of 60 Chinese late learners of English. Flege & Liu (2001) found that LOR was a crucial factor regarding L2 outcomes for late learners, but only if they needed to use English regularly. It remains unclear if this was due to the quality of the input (from native speakers), the quantity, or a combination of both. Further complicating the issue, L2 input is likely associated with other factors, such as L2 feedback, motivation, and attention, to name just a few.

In comparison with input, the role of L2 use in the production/perception of adult learners has received less attention. There is, however, an illustrative example in a series of foreign accent studies conducted by Flege and colleagues. Flege et al. (1995) found that the earlier the age of learning (AOL), the more native-like the participants' production/perception. In an unpublished follow-up to this study, Flege & MacKay revisited the learners from the original investigation 10 years later (as cited in Flege, 2012).

The researchers found that increased use of English was associated with more native-like production when compared with the 1992 data. A decrease in English usage resulted in no change in production accuracy. In a separate unpublished longitudinal study, Flege found that after a period of 5 years in the U.S., foreign accent ratings of Spanish-speaking late learners of English showed no improvement (as cited in Flege, 2012). However, post-hoc scrutiny of the data compared the 3 “worst” (more foreign accented) learners with the 3 “best” (least foreign accented) learners and showed that the “best” learners reported using English more. Specifically, they used English in contexts where it could be considered optional (i.e., in conversations with friends), suggesting that L2 use in extensive contexts may foment L2 phonological development.

In sum, the role of input and use in adult language learning has not been a primary object of focus in the SLA literature despite the crucial status these factors are given in L1 acquisition. This is likely the result of both variables being overshadowed by research on age effects, coupled with the fact that they are difficult to control and there is no straightforward method for quantifying them. Language immersion provides an ideal point of comparison for studying use and input because learner access to native speakers is high and the target language is likely used often. Moreover, this learning context opens the door to studying L2 phonological acquisition during the initial stages of learning. The present study contributes to this literature by examining adults in a domestic immersion program in which L2 use is maximized and L2 input is rampant.

Production/perception interface in L2 learning

The relationship between production and perception is assumed to be non-controversial, as there are numerous studies demonstrating the relationship between the two modalities (i.e. Flege & Eefting, 1988; Williams, 1979; Flege et al., 1999, among many others). However, there are discrepancies in the literature regarding which of the two modalities is the driving force in L2 learning.

On one hand, a long line of research supports the claim that perception of a novel phonetic segment precedes its production (Williams, 1979; Borden et al., 1983; Neufeld, 1988; Barry, 1989; Grasseger, 1991; Flege, 1993; Rochet, 1995; Llisterri, 1995; Flege et al., 1997; Leather, 1999). For instance, Flege et al. (1997) explored the production/perception of /i/-/ɪ/ in Spanish-speaking learners of English and found that a subset of the participants perceived the vowels similarly to a group of native controls. Crucially, only a few of the native-like perceivers were able to produce the /i/-/ɪ/ contrast accurately. Flege et al. (1997) took these findings as evidence that accurate perception must necessarily precede accurate production. Further support for the perception-first hypothesis comes from training studies. For example, some studies on perceptual training have lead to more accurate production (See Rochet, 1995; Bradlow et al., 1997, 1999).

On the other hand, there are also investigations that cast doubt on the perception-first claim (Goto, 1971; Caramazza et al., 1973; Sheldon & Strange, 1982; Mack, 1989; Mathews, 1997; Leather, 1997; Wang, 2002; Kissling, 2014). For instance, Sheldon & Strange (1982) examined the production/perception of American English /l/-/r/ in Japanese learners and found that they demonstrated more accuracy in production than in discrimination. The authors contend that production accuracy may precede perceptual accuracy for a given segment due to pedagogical reasons, as some of the participants received instruction regarding the contrast in question based on articulatory parameters (as opposed to auditory cues). Mack (1989) obtained similar results in English/French bilinguals and hypothesized that there might be greater consequences associated with mispronouncing a segment versus misperceiving it. Alternatively, the abundance of viable acoustic cues in the speech signal might make sufficient perception possible without monolingual-like mastery of any single cue.

One method of exploring the nature of the production/perception relationship is via longitudinal data. In a semi-longitudinal analysis of Spanish /p/-/b/, Zampini (1998) and Zampini & Green (2001) tested intermediate/advanced learners on 3 separate occasions during the course of an academic semester. These studies found that perception changes occurred after production changes, though one cannot reach a definitive conclusion from these data regarding the nature of the relationship between the two modalities because the participants in question were fairly advanced and, at the time of testing, already differed from monolinguals in their perception of Spanish stops. Longitudinal data from absolute beginners are necessary to determine which of the two modalities occurs first in L2 phonological development.

L2 speech models and phonetic category development

There are numerous L2 models used to posit hypotheses that can account for the difficulties encountered by L2 learners. Relevant to the present work are the Speech Learning Model (SLM, Flege, 1995) and the Linguistic Perception Model (L2LP, Escudero & Boersma, 2004; Escudero, 2005, 2009). The SLM maintains that phonetic similarity or perceived equivalence can predict L2 difficulties. According to the SLM, the human ability to learn novel sounds is maintained throughout life. Representations of the L2 sounds being acquired are stored in long-term memory and share a common phonological space with the L1. The SLM proposes that bilinguals aim to maintain L1 and L2 phonetic categories separate. This implies that L1 to L2, as well as L2 to L1, interactions occur, with each language having some influence on the other. Formation of novel L2 categories becomes more difficult—but not impossible—as the L1 sound system develops. If L1 and L2 sounds are too perceptually similar, category formation is hindered because learners perceive the L2 sound as being equivalent to the L1 category.

For its part, the L2LP also maintains that the human ability to learn novel speech sounds remains active throughout life and that L2 difficulties are accounted for via phonetic similarities, differences, or perceived equivalences to native contrasts. The L2LP posits two distinct situations for learning L2 contrasts. A contrast can be considered novel (“new scenario”) to the learner, in which case the model contends that new categories must be formed. Conversely, a contrast can be familiar ("similar scenario"), and (s)he must then reset the boundary for the contrast via a comparison module, the Gradual Learning Algorithm (GLA).

Under the L2LP, a new L2 grammar is created via *full copying/full access*, and is developed independently of the L1 perception grammar. Thus L2 learners have the ability to become native-like in their perception of the L2 without affecting the L1 perception grammar. This notion is at odds with other popular speech models, namely the SLM, which proposes that L1 and L2 categories share the same phonetic space and therefore L2 development is hypothesized to occur simultaneously with changes in L1 categories. The L2LP, on the other hand, proposes that L2 learning occurs as a result of resolving two problems. The first problem deals with the fact that the perception grammar must change, or adjust, in order to manage input from the L2, and is, thus, a perceptual problem. The second problem, representational in nature, requires that new L2 categories be created whenever L1 categories cannot be used. Similar contrasts are hypothesized to be easier to learn because listeners can simply reuse L1 categories. Phonetic differences are accounted for via boundary adjustments. Adjustments to the L2 perception grammar also occur through the GLA. While the SLM and the L2LP clearly differ regarding how phonetic category formation occurs, both models assume an underlying relationship between production and perception. According to Flege, L2 phonetic segments are "[. . .] produced only as accurately as they are perceived" (Flege, 2003, p. 25). The L2LP formalizes this relationship by positing a specific perceptual grammar that works in conjunction with a production grammar. Escudero (2005) maintains that the model is best utilized with longitudinal data collected from beginning language learners as they progress over time. Accordingly, the present work examines the production/perception of pure beginners during the initial stages of L2 learning.

Spanish and English stop contrasts

The voicing contrasts of Spanish and English stops serve as a proxy for analyzing the production/perception relationship and the roles of input and use during early L2 phonological development in the present study. Stops can vary based on their point of articulation and voicing. Most, but not all languages, have voiceless stops, and, of these languages, many have voicing contrasts. While some languages have three-way contrasts for stops (e.g., Thai), most have a two-way voicing contrast. Two-way contrast languages arbitrarily fall into one of two categories: true voicing or aspirating languages. English and Spanish are languages of the two-contrast variety. English is an aspirating language; Spanish is a true voicing language.

Importantly, these languages share the same two-way contrasts—phonologically voiced versus voiceless phones at bilabial, coronal, and velar place—, but differ in where the acoustic boundary lies between them based on voice onset time (VOT, Lisker & Abramson, 1964). VOT refers to the duration of the time interval between the release of the stop burst and the onset of modal voicing. English, an aspirating language, contrasts /b, d, g/ with /p, t, k/ through phonetically voiceless stops with short-lag (positive) VOT and phonetically aspirated stops with long-lag (positive) VOT, respectively.[1] Spanish, a true voicing language, contrasts the same phonological segments through phonetically prevoiced stops with lead (negative) VOT and phonetically voiceless stops with short-lag (positive) VOT (Lisker & Abramson, 1964).

These differences may appear trivial, however, they have important implications with regard to speech production/perception. L2 learners are likely to produce and perceive L2 stops using the acoustic boundaries of their L1 (Flege, 1987b). As productions in the L2 deviate from native values, they are more likely to be perceived as sounding foreign. Thus,

[1] In English phonologically voiced stops can surface partially or fully voiced in prosodically weak positions in connected speech (Davidson, 2018). This is less common in utterance initial position.

accurately producing the stop contrasts of a foreign language can significantly improve foreign accent ratings (Sundara, Polka, & Baum, 2006). For example, if a native English speaker learning Spanish produces the long-lag VOT associated with English /p/ (i.e. aspiration, [p^h]) when saying the Spanish word *papel* ("paper")—which is realized as short-lag [p]—, the production would sound foreign to the native Spanish speaker (Elliott, 1997; González-Bueno, 1997; Sundara, Polka, & Baum, 2006). The issue is also complex for the L2 learner of Spanish regarding perception. Given that both languages have short-lag VOT in phonologically distinct segments, the L2 learner can—and often does—mistakenly "hear" an English /b/ category when the utterance, in reality, contained Spanish /p/. This could be the difference between "hearing" *peso* ("I weigh") versus *beso* ("I kiss").

The task of English-speaking L2 learners of Spanish is to associate short-lag VOT with Spanish voiceless /p, t, k/, and create a new category for lead VOT associated with voiced /b, d, g/ in order to produce/perceive Spanish stops accurately. Recall that both the SLM and the L2LP maintain that L2 learners develop new phonetic categories. Under the L2LP account, learning the voicing contrasts of Spanish involves boundary resetting (i.e., similar scenario) and is hypothesized to be relatively easy for the learner. The SLM, on the other hand, would posit that Spanish stops might pose a challenge to the L2 learner of Spanish because of the phonetic similarity between the corresponding English and Spanish phones. Accordingly researchers have attempted to demonstrate phonetic category formation by examining the acoustic properties of L2 speech and comparing them to those of native speakers.

For example, Flege & Eefting (1987) examined Spanish speakers' production of word initial English stops and found that they produced the English segments with aspiration. However, the acoustic analyses showed that the VOT values were longer than in Spanish words, but not as long as the VOT values of a monolingual English control group. Flege & Eefting (1987) concluded that the Spanish/English bilinguals did indeed develop L2 phonetic categories, but their productions were not "authentic" because they still differed from those of the monolingual controls. They ascribed these differences to the nature of the input their participants had received, i.e., Spanish-accented English. The aforementioned study suggests that learners are capable of acquiring new phonetic categories, though the nature of these categories may differ from those of monolingual speakers.

Previous research on the acquisition of Spanish stops has included traditional classroom learners (Zampini, 1998; Zampini & Green, 2001; López, 2012; González López & Counselman, 2013; Nagle, 2017), as well as study abroad (SA) immersion learners (Stevens, 2001; Díaz-Campos, 2006; Nagle et al., 2016). For instance, a study by Stevens (2001) compared the production of /p, t, k/ of a group of North American English speakers participating in an immersion program in Spain to a group of traditional classroom room learners. Stevens (2001) found that the SA group learned to reduce aspiration in their Spanish productions, and the traditional classroom learners continued producing Spanish voiceless stops with long-lag VOT. Moreover, Stevens (2001) noted a linear relationship between the length of stay in the foreign country and production accuracy and suggested that the extensive use of Spanish likely accounted for the SA groups' phonetic category development. In a similar vein, Díaz-Campos (2006) found that SA learners showed greater improvements than a traditional classroom control group in the production of /p, t, k/ when their speech was analyzed in a conversational style. In more formal settings, acoustic analyses showed no differences between the same groups. There is also evidence of category formation in traditional classroom learners. For instance, in a semi-longitudinal study of Spanish bilabial stop production, Zampini (1998) showed that a group of North American intermediate/advanced late learners produced Spanish voiceless stops with shorter VOT than their English stops, but longer VOT than native Spanish speakers (see also Zampini & Green, 2001, for other cues). The learners did not incorporate prevoicing into their productions of /b/ by the end of the semester.

Summarizing, the literature on the adult acquisition of Spanish stops has focused on the traditional classroom and immersion contexts. The findings suggest that adults can form new phonetic categories for Spanish voiceless stops—especially in a learning context that facilitates L2 use and provides ample native input—, but voiced stops may take longer to develop. There is a gap in the SLA literature regarding the initial stages of adult L2 phonological acquisition, and in particular regarding Spanish stops.

The present study

To investigate phonetic category development during the early stages of L2 learning, the present work tracked the production/perception of 10 adult native English speakers who participated in a Spanish immersion program. The goals of this study were to (1) explore phonetic category development related to the fine-phonetic detail of Spanish stop voicing contrasts, and (2) determine whether L2 phonological development is driven by production or perception. Furthermore, this research contributes to the L2 literature regarding language input and use by providing semi-longitudinal data from beginning adult learners in a context in which L2 use and input were maximized. Conversely, L1 use was held to a minimum.

General method

Questionnaires

The learners completed a language background questionnaire and a weekly progress questionnaire. The former excluded participants who had experience learning Spanish or other languages. The questions inquired about time spent studying an L2, living in a foreign country, if they had family members who spoke a language other than English, and provided an overall assessment of their Spanish. The second questionnaire was a self-report assessment aimed at quantifying information related to weekly language use and input, along with other measures. The questionnaire asked about the participants' use of Spanish/English, their self-reported speaking abilities, listening/comprehension abilities, overall abilities in Spanish, and if they felt their Spanish had improved. The learners completed this questionnaire every Sunday before participating in the experimental tasks. Thus responses each week referred to what they had done during the previous week. The questions related to their use of Spanish provided a percent estimate of time spent speaking Spanish with native speakers, and non-native speakers with more, less or equal proficiency. A third questionnaire, the Bilingual Language Profile (BLP, Birdsong et al., 2012), was administered to a bilingual control group. The BLP provided a measure of language dominance by calculating scores for 4 modules: language history, language use, language proficiency, and language attitudes. The summed scores from each module provide a value of language dominance that ranges from -218 to 218. A score near either extreme indicates dominance in one of the two languages. Negative scores were associated with dominance in English, and positive scores were associated with dominance in Spanish. A score near 0 indicated balanced bilingualism.

Participants

The present study included 20 people who were divided into 2 groups. The first group consisted of 10 adult L2 learners of Spanish whose native language was English. The second was a control group comprised of 10 simultaneous Spanish-English bilinguals.[2]

[2] The use of bilingual control groups, as opposed to the more traditional practice of using monolingual speakers, is part of a recent trend in L2 phonetic research (see Sakai, 2018). Numerous investigations show even early bilinguals with ample L2 proficiency tend to differ from monolinguals in production, perception, and lexical processing (i.e., Pallier et al., 1997; Sebastián-Gallés & Soto-Faraco, 1999; Sebastián-Gallés et al., 2005; Sebastián-Gallés, 2006, among others). These differences are often ascribed to cross-linguistic interactions, which, simply put, may well be a part of bilingualism, independent of age of acquisition and L1/L2 proficiency. The present work takes the position that individuals undertake the endeavor of language learning with the goal of becoming bilingual and not to replace their native language. Therefore a more fair and useful assessment of their progress can be achieved by comparing their abilities to those of a population (bilinguals) that represents their end goal (bilingualism).

Late learners. The learners of the present study were students in a domestic immersion language program at Middlebury College. The defining characteristic of the program is the Language Pledge, a formal agreement the students signed by which they promised to use only the target language (in this case Spanish) for 7 weeks. Failure to comply with the pledge can result in expulsion. Students lived in the residence halls on the campus with other students and professors, and they attended class for 4 hours in the morning and participated in co-curricular activities in the afternoon. The program is designed with the intention of creating an experience comparable to living abroad, though it is considered intense immersion due to the pledge and the seriousness of the students and the curriculum. The classes employed a communicative focus with instruction entirely in Spanish.

Information from the background questionnaire was used to select 10 participants (4 males, 6 females) who reported no prior experience with any other languages. Individuals who had completed a semester or more of foreign language study or had spent time living in a foreign country were excluded. The late learners were 18 years old or older ($\bar{x} = 23.70$; SD = 5.27), and considered absolute beginners. This assertion was confirmed via placement testing and an interview with two faculty members of the immersion program.

Table 1 displays a summary of the self-report data obtained from the weekly assessment questionnaire. The learner group used their L1, English, minimally, and their L2, Spanish, almost exclusively. The L2 input to which the learner group was exposed was provided mainly by native speakers or non-native speakers with higher levels of proficiency in Spanish. Furthermore, the learners believed their listening, speaking, and overall abilities in Spanish improved over the course of the program.

++ INSERT TABLE 1 ABOUT HERE ++

Simultaneous (native) bilinguals. Ten simultaneous (i.e., native) bilinguals participated in the present study. These native Spanish and English speakers served primarily as a control group with which the production/perception of the learner group was compared. Bilingual participants reported speaking both English and Spanish for as long as they could remember, and that their parents were also bilingual. Moreover, bilingual participants stated they used both languages on a daily basis with friends and family. The mean language dominance score of the bilingual group was -2.76 (SD = 38.93), suggesting balanced bilingualism according to the BLP.

Overview of procedures

The learners completed four distinct tasks: two related to their production in Spanish and two related to their perception in Spanish. The present study reports one of the production tasks and one of the perception tasks. The purpose of the tasks was to provide data measuring their progress learning Spanish bilabial stops. During the initial session, the learners completed the first iteration of the assessment questionnaire, a delayed repetition production task, and a two-alternative forced-choice perception task (2AFC). From this point forward until the final week of the program, the learners completed the same tasks with the exception that the delayed repetition production task was replaced with a reading production task. Experimental sessions took place every Sunday with the exception of the last week, which included multiple sessions and two tasks not reported here. Table 2 presents an overview of the learners' participation.

++ INSERT TABLE 2 ABOUT HERE ++

Bilingual participants were recruited from a university in the Southwestern United States. Their participation included two days of testing at a speech science laboratory with a minimum of 24 hours between sessions. On the first day they completed the BLP questionnaire, the 2AFC task, and the reading task. On the second day they completed tasks that are not reported here. The following sections report the results of the production task, proceeded by the perception task, and, finally, a comparison of the two modalities.

Longitudinal development of L2 bilabial stop production

The first task examined the ongoing development of Spanish bilabial stop production in adult L2 learners with a special focus on how the realization of stop voicing changed with increased L2 exposure.

Method

Materials.

Target phrases.

Participants repeated a series of words containing Spanish stops, /p, t, k, b, d, g/, in utterance initial position. A total of 30 nonce words—5 for each stop segment—were embedded in the carrier phrase "______ es la palabra" (*Eng.* "______ is the word"). The syllable structure was CV.CV with primary stress falling on the initial syllable. Stops were followed by one of the 5 Spanish vowels, /i, e, a, o, u/, and the consonant /k/. The final vowel was always the same as the first (i.e., /ˈ_a.ka/, /ˈ_e.ke/, /ˈ_i.ki/, etc.). There were 6 stops (/p, t, k, b, d, g/) × 5 vowels (/i, e, a, o, u/) = 30 items. The present work focused on the bilabials, /p, b/. Target words were interspersed amongst 20 distactors used in another task.

Auditory stimuli.

A 29 year old native female Spanish speaker from Cádiz, Spain, provided the audio stimuli presented in the delayed repetition portion of the task. A Shure SM10A dynamic head-mounted microphone recorded the items. A Sound Devices MM-1 pre-amplifier boosted the signal and sent it to a laptop computer were it was recorded using Praat at a 44.1 kHz sample rate with 16-bit quantization (Boersma & Weenink, 2018). The recording took place in a sound attenuated booth in a phonetics laboratory at a university in the U.S. southwest.

Procedure.

Recordings.

The learners were recorded in a quiet classroom on site at Middlebury College. A Shure SM10A dynamic head-mounted microphone captured the participants' productions. A Sound Devices MM-1 pre-amplifier passed the signal to a Marantz PMD661 MKII Handheld Solid State broadcast recorder. Recordings were sampled at 44.1 kHz with 16-bit quantization. The same setup used to record the auditory stimuli served to record the productions of the bilinguals.

Acoustic analysis.

Participants' productions were segmented in Praat using synchronized waveform and spectrographic displays to hand-mark the onset of voicing and the burst for each stop. Voicing onset was the first periodic pattern found in the waveform. The criterion for bursts was the onset of broad-band sudden noise in the spectrogram. A Praat script automatically extracted VOT, which was calculated as the difference (in ms) between the aforementioned acoustic landmarks (i.e. onset of modal voicing and the burst). Figure 1 illustrates the segmenting procedures.

++ INSERT FIGURE 1 ABOUT HERE ++

The initial experimental session took place on the second or third day of the immersion program. Thereafter, data collection took place every Sunday until the end of the program for a total of 8 experimental sessions. The final session took place on a Wednesday, which was 3 days after the seventh session and 3 days before the program ended. PsychoPy2 (Peirce, 2008) presented the stimuli randomly. The first session employed a delayed shadowing technique. The stimuli were presented aurally via the audio recordings of the native Spanish speaker. Participants listened and repeated the 50 items embedded in the carrier phrase. Subsequent experimental sessions utilized a reading task.

In these cases, the stimuli appeared on a computer screen and participants read them aloud.[3] After saying the stimuli aloud, participants pressed a button on a keypad to advance to the next item. Each learner provided the dataset with 240 bilabial stops (2 stops × 5 vowels × 3 repetitions × 8 sessions). Thus, a total of 2,400 stop tokens were collected from the learner group (240 tokens × 10 participants = 2,400). The bilingual controls only completed one session of the reading task, and provided a total of 300 utterance initial stops (2 stops × 5 vowels × 3 repetitions × 10 participants). The task took approximately 10 minutes.

Statistical analyses. Data from the production task were analyzed using a series of generalized linear mixed effects models (GLMM). Specifically, there were 3 analyses. The first analysis aimed to see if learners' production of Spanish bilabial stops changed by end of the immersion program, and how self-reported measures of input and use affected their progress. In this model change in VOT in standardized units (ΔZ_{VOT}) was the criterion. To calculate ΔZ_{VOT}, raw VOT values were converted to z-scores (i.e., standardized) as a function of voicing. For each participant, the difference in standardized VOT from the start of the program and the end of the program was calculated. The result was a value indicating whether the VOT of learners' bilabial productions had reduced (negative ΔZ_{VOT}), increased (positive ΔZ_{VOT}), or remained the same (ΔZ_{VOT} near 0) after the program. Fixed effects were averaged self-report assessments of input (% of input from the following sources: (1) native, (2) non-native, (3) non-native with higher proficiency, (4) non-native with lower proficiency, and (5) non-native with equal proficiency) and use (% time speaking Spanish, English), with by-subject random effects on all continuous predictors and item repetitions. The fixed effects were standardized, thus all

[3] Two separate techniques were used in the production task (delayed repetition during week 0, reading during the remaining weeks) in order to avoid intervocalic 't' being realized as [ɾ] in a subset of the distractors that served as stimuli for another experiment. Participants did not produce intervocalic 't' as [ɾ] in any of the experimental sessions.

continuous predictors had a mean of 0 and a standard deviation of 1.

The purpose of the second analysis was to determine the amount of exposure time necessary for the learners' production of Spanish bilabial stops to change. Production data for /b/ and /p/ were fit separately. VOT was the criterion and *exposure time* (day 0, day 7, day 21, day 28, day 35, day 42) was a fixed effect. The random effects structure included by-subject and by-item intercepts with slopes for exposure time (for the subject effect) and item repetitions (for the items effect). Exposure time was dummy coded with day 0 set as the reference level. Thus all tests of simple effects compared VOT values after a given amount of exposure (i.e., day 7, day 14, etc.) to the baseline.

The final analysis directly compared the learners' production of Spanish bilabial stops at the end of the immersion program (day 42) with that of the bilingual control group. VOT was again the criterion and the regressors were *group* (learner, native bilingual) and *voicing* (voiced, voiceless). The model included by-subject and by-items intercepts with random slopes for voicing and item repetitions. The fixed effects were again dummy coded with native bilinguals' voiceless stops set as the reference levels.

For all models, visual inspection of Q-Q plots and plots of residuals against fitted values were used to check for normality of the residuals. Unless noted otherwise, statistical significance of main effects and higher order variables was assessed using hierarchical partitioning of the variance via nested model comparisons. Marginal R^2 and conditional R^2 provided an indication of goodness-of-fit for each model (Nakagawa & Schielzeth, 2013). Marginal R^2 specified a measure of variance explained without random effects and conditional R^2 included them.

Results

The role of input and use in bilabial stop production.

++ INSERT FIGURE 2 ABOUT HERE ++

Panels (a) and (b) of Figure 2 plot the VOT production data as a function of days of exposure, voicing and group (learner, native bilingual). The first analysis examined ΔZ_{VOT} as a function of self-reported input and use factors. The omnibus model tested the hypothesis that ΔZ_{VOT} was significantly different from 0. The intercept estimate was -0.74 $\pm$ 0.10 standard errors (CI low = -0.93; CI high = -0.56; t = -7.8; $p < 0.001$). The negative estimate indicates a net decrease in VOT for bilabial stop production. Back-transforming to raw values showed that voiced stops had an average VOT of 13.11 ms and lowered by -38.49 ms by the end of the program. Voiceless stops were 19.2 ms higher at baseline ($\bar{x}$ = 32.31 ms) and decreased by approximately -10.78 ms by the end of the program. Continuous input and use predictors were included in the model using forward selection and only retained if they significantly contributed to model fit. There was only one main effect: self-reported % of English use ($\chi^2(1) = 4.98$; $p < 0.027$). Specifically, a 1-unit increase of *Z-English use* was associated with an increase in ΔZ_{VOT} of 0.23 $\pm$ 0.10 standard errors (CI low = 0.03; CI high = 0.43; t = 2.29; $p < 0.03$). Thus learners who self-reported higher overall English use showed smaller changes in VOT (See Figure 3). The model of best fit included random effects (R^2m = 0.03; R^2c = 0.50).

++ INSERT FIGURE 3 ABOUT HERE ++

Change in bilabial stop production over time.

Voiceless bilabial stops.

The voiceless bilabial data were best fit using a maximal error term (R^2m = 0.07; R^2c = 0.66). There was a main effect of session ($\chi^2(6) = 27$; $p < 0.001$). VOT values had lowered by -8.86 ms $\pm$ 3.41 standard errors (CI low = -15.55; CI high = -2.18; t = -2.6; $p < 0.02$) after 21 days of exposure. The average VOT difference from the baseline value was approximately 10 ms from this point forward; however, there was appreciable variability, as seen in the standard errors of the parameter estimates for each testing session. Table 3 provides the complete model output and Figure 2 shows the distributions of the data at each testing session.

++ INSERT TABLE 3 ABOUT HERE ++

Voiced bilabial stops.

The analysis showed that the voiced bilabial stop data were also best fit using a maximal error term ($R^2m = 0.16$; $R^2c = 0.67$). There was a main effect of session ($\chi^2(6) = 29.95$; $p < 0.001$). VOT was significantly lower than the week 0 initial state after 21 days of exposure (CI low = -49.89; CI high = -13.13; t = -3.36; $p < 0.001$), and for each session thereafter. Thus, learner VOT values for voiced bilabials decreased after three weeks in the program. Table 4 displays the percentage of prevoiced /b/ as a function of exposure time. Crucially, all learners produced prevoiced stops at least some of the time, and by the conclusion of the program approximately half of the productions included lead VOT. The average VOT of the prevoiced stops was lower ($\bar{x} = -75.3, SD = 35.3$), suggesting production of /b/ was inconsistent. The complete model output is displayed in Table 5 and density ridgeline plots of the distributions for each session are available in Figure 2.

++ INSERT TABLE 4 ABOUT HERE ++

++ INSERT TABLE 5 ABOUT HERE ++

Comparison with bilinguals. The data were best fit when including the random effects structure ($R^2m = 0.54$; $R^2c = 0.73$). There was no effect of *group* ($\chi^2(1) = 1.26$; $p > 0.05$), but there was an effect of *voicing* ($\chi^2(1) = 23.28$; $p < 0.001$), as well as an interaction between the two factors ($\chi^2(1) = 11.92$; $p < 0.002$). The learner groups' voiceless stops had a mean VOT value of 22.80 ms, approximately 4.77 ± 3.3 standard errors higher than the controls, a difference that was not statistically significant (CI low =

-1.7; CI high = 11.25; $t = 1.44$; $p > 0.05$). The learner groups' voiced stops, on the other hand, differed from those of the control group by 53.87 ± 13.35 standard errors (CI low = 27.7; CI high = 80.04; $t = 4.03$; $p < 0.001$). As shown in the previous analysis, the productions of /b/ that were indeed prevoiced were closer to the bilingual range. The distributions for voiceless and voiced stops of the bilinguals are displayed in the final ridgelines of panels (a) and (b) of Figure 2.

Interim discussion

The first task investigated bilabial stop production in late learners of Spanish. The task was designed to determine (1) if the learner group improved its production of bilabial stops after a 7-week immersion program, and the extent to which input and use factors modulated this improvement, (2) how much exposure was necessary for observable phonetic category development to take place, and (3) how the learners production compared to a group of simultaneous bilinguals upon completion of the immersion program.

Upon comparing stop production from the baseline initial state and the final testing session after 42 days of exposure to Spanish, the results of the first analysis suggested that overall the learners reduced VOT for bilabial stops. Specifically, by the end of the program they reduced aspiration for the voiceless stops and began to incorporate prevoicing into their voiced stops. These changes were partially modulated by self-reported use of English, such that participants who reported higher overall use of English during the program showed less improvement in bilabial stop production. The second analysis found that the learners' production boundaries began to shift after 21 days of exposure to Spanish. That is, after the third week in the program, both /p/ and /b/ had lower VOT values, and continued to decrease throughout the remainder of the program. The third analysis compared the learners' stop production in the final week of the program to that of the bilingual control group. Although the learners reduced VOT for both stop segments, they clearly differed from the bilinguals regarding the voiced segment /b/, despite the fact that this segment appeared to show a larger change (in ms) by the end of the program. An analysis of the proportion of prevoiced stops showed that all learners included lead-vot some of the time, and, when they did, the values were much closer to the bilingual range, suggesting /b/ production was particularly unstable. The voiceless stop was produced within the range of native values for VOT. In sum, the first task showed that (i) the learners did improve their stop production in a 7-week immersion program, (ii) use of English affected production gains, and, finally, (iii) evidence of phonetic category development was observable after 21 days of exposure.

L2 perception of bilabial stops

The second task examined the ongoing development of Spanish stop perception in adult L2 learners with a special focus on how stop voicing identification changed with increased L2 exposure.

Method

Materials. In order to create a VOT continuum a twenty-three year old female Spanish/English simultaneous bilingual from the Southwestern U.S. provided natural productions of the bisyllabic words "bata" (Eng. *robe*) and "pata" (Eng. *paw*), each of which contain stops in utterance initial position. An AKG C520 condenser microphone was used to record the utterances. A Sound Devices USBPre 2 audio interface digitized the signal at 44.1 kHz and 16 bit quantization. The digitized signal was sent to a laptop computer and recorded using Praat (Boersma & Weenink, 2018). The best token of [p^h]—one in which there were no clicks or extraneous noise—was selected for resynthesis. For the stimuli with positive VOT, Praat manipulated the duration of the aspirated portion of the stop via the Time-Domain Pitch-Synchronous-Overlap and Add algorithm (TD-PSOLA). For the stimuli with negative VOT, periods of prevoicing were pasted into the signal at zero-crossings before the release of the stop. The prevoiced portions were taken from phonetically voiced stop productions of the aforementioned simultaneous bilingual. The result was a VOT continuum ranging from -60 to 60 ms in 10 ms increments. Finally, the stimuli were normalized for peak intensity.

Procedure. Upon completing the questionnaires, the learners participated in the 2AFC task. The initial experimental session took place on the second or third day of the immersion program. Subsequent data collection occurred every Sunday for the remainder of the program for a total of eight experimental sessions. The final experimental session took place on a Wednesday, which was five days after the seventh session and three days before the program ended. PsychoPy2 (Peirce, 2008) presented the stimuli described above via a Macbook Pro. The program produced the audio stimuli at the same time that the orthographic labels “ba” and “pa” appeared on the left-hand and right-hand sides of the screen, respectively. The participants then determined whether they had heard “ba” or “pa” by pressing the appropriate button on a DirectIN Rotary Controller. A red cross appeared in the middle of the screen between trials, indicating a new trial was about to begin. There was no set time limit for each trial; however, participants were instructed to respond as quickly and as accurately as possible. The program presented one stimulus per trial in ten randomized blocks (13 stimuli $\times$ 10 blocks = 130 tokens) with the inter-stimulus interval set at 500 ms. The participants finished the task in approximately 8 minutes. The task was administered in 8 separate sessions, once per week until the conclusion of the program.

Statistical analyses. Data from the perception task were analyzed in `R` (R Core Team, 2017) and can be separated into two principle analyses, one focused on the learners perceptual behavior over time, and the other was concerned with how the learners compared with bilingual controls at the offset of the immersion program. First, a series of models were fit to examine the learners’ perceptual identification as exposure to Spanish increased. Due to the categorical nature of the participants’ responses (i.e. “ba” or “pa”), the data were analyzed using a GLMM with a binomially distributed error term and logit linking function. The omnibus model was fit with *VOT* and *exposure time* as continuous fixed effects, and self-report assessments of input (% of input from the following sources: (1) native, (2) non-native, (3) non-native with higher proficiency, (4) non-native with lower proficiency, and (5) non-native with equal proficiency) and use (% time speaking Spanish, English) variables were included using forward selection. Causal priority was given to *exposure time*. All predictors were standardized such that their mean value was 0. The random effects structure included a scalar random effect for each subject with random slopes for *exposure time* and *VOT*.

Next, a second series of models was fit to examine how the learners' perceptual boundaries of the resynthesized continuum shifted as exposure increased. The random effects output from the aforementioned omnibus model was utilized to determine the 50% boundary crossover point (CO) for each participant at each testing session. The CO for the boundary between voiced and voiceless stops was calculated using the `cross_over` function of the package `lingStuff` (Casillas, 2018). This function calculated the perceptual boundary using the following formula:

$$CO = \frac{\beta_0}{\beta_{VOT}} \times -1 \tag{1}$$

where each by-subject intercept (β_0) is divided by the estimated by-subject slope for the effect VOT (β_{VOT}) and multiplied by -1. The CO point values were standardized and served as the dependent variable in subsequent analyses. In order to assess perceptual boundary shifts over time, the CO data were analyzed using a GLMM with *exposure time* as the dummy coded fixed effect. Day 0 was set as the reference level, thus the omnibus model provided parameter estimates of the change in CO values as time progressed with regard to the baseline perceptual boundary. The random effects structure included by-subject intercepts with a random slope for exposure time.

The final analysis compared the learners' perception of the resynthesized continuum at the end of the program with that of the bilingual control group. This analysis included three models. The first model fit the identification response data ("ba", "pa") as a function of group (learners on day 47, bilingual controls) and VOT. The second model scrutinized the perceptual boundary (CO) data as a function of group. The final model examined contrast coefficient slopes (CCS) as a function of group. Contrast coefficient slopes in the logistic space were calculated for the corresponding sigmoidal curves in the probability space. The contrast coefficient slope gives a measure of "crispness" between phoneme boundaries (Morrison, 2007)[4], and were derived for each participant using the parameter estimate for VOT from the random effects output of the omnibus model and the following equation:

$$CCS = \beta_{VOT} \times .25 \quad (2)$$

where the estimated by-subject slope for the effect VOT (β_{VOT}) was multiplied by .25. For the three aforementioned models *group* was deviation coded (-0.5, 0.5), thus the model parameter estimates provide an assessment of effect size. All mixed effects models were fit using the `R` package `lme4` (Bates et al., 2015). Main effects and higher order interactions were assessed using hierarchical partitioning of the variance via nested model comparisons. Visual inspection of Q-Q plots and plots of residuals against fitted values were used to check for normality of the residuals for linear models fit using Gaussian distributions. Marginal R^2 and conditional R^2 again provided an indication of goodness-of-fit for each model (Nakagawa & Schielzeth, 2013).

[4] Speakers are believed to have "crisp" boundaries between native contrasts. When learning a new contrast, L2 speakers often have "fuzzier" boundaries, represented by shallower slopes. See Morrison (2007) for discussion on this topic.

Results

Input, use, and perceptual categorization over time.

Input, use, and perceptual categorization over time. The GLMM yielded a main effect of *VOT* ($\chi^2(1) = 24.88$; $p < 0.001$), *exposure time* ($\chi^2(1) = 4.92$; $p < 0.028$), as well as a *VOT* x *exposure time* interaction ($\chi^2(1) = 4.78$; $p < 0.03$). The model containing the higher order interaction was retained. There were no main effects nor interactions related to the input and use predictors. The model output revealed that the log odds of responding "voiceless" increased by 5.25 ± 0.4 standard errors as VOT increased (CI low = 4.46; CI high = 6.03; $z = 13.16$; $p < 0.001$). Overall, there was a change in the log odds of "voiceless" responses of 0.19 ± 0.09 standard errors as a function of exposure time (CI low = 0.01; CI high = 0.37; $z = 2.11$; $p < 0.05$). Moreover, the VOT x exposure time interaction corresponded with a slope adjustment of 0.25 ± 0.11 standard errors (CI low = 0.03; CI high = 0.47; $z = 2.21$; $p < 0.04$), suggesting that the probability of responding "voiceless" at the baseline VOT was higher as exposure to Spanish increased. Figure 4 plots predicted "voiceless" responses as modulated by VOT at each testing session. One can observe a sigmoid function that appears to phase shift to the left over time.

++ INSERT FIGURE 4 ABOUT HERE ++

The analysis of the boundary crossover point data revealed a main effect of session ($\chi^2(7) = 18.3$; $p < 0.012$). Specifically, the crossover point was significantly different from the week 0 baseline values after 14 days of exposure. At this point, the boundary had decreased by -0.66 standardized units ± 0.29 standard errors (CI low = -1.23; CI high = -0.09; z = -2.29; $p < 0.04$). The remaining sessions also showed significantly lower boundary crossover points with the exception of day 35 (on the sixth experimental session). The complete model output is shown in Table 6.

++ INSERT TABLE 6 ABOUT HERE ++

Comparison with bilinguals. Figure 5 plots the results from the three models comparing the learners to the bilingual control group. Concretely, panel (a) displays the sigmoidal curves in the probability space along with the group CO points, and panel (b) plots the corresponding contrast coefficient slopes in the logistic space (b). The GLMM comparing the learners with the bilingual control group was best fit when including the random effects structure ($R^2m = 0.88$, $R^2c = 0.95$). The model yielded a main effect of *VOT* ($\chi^2(1) = 39.51$; $p < 0.001$). For both groups, a 10 ms increase in VOT was associated with a 0.2 ± 0.02 standard errors increase in the log odds of responding voiceless (CI low = 0.16; CI high = 0.23; $z = 11.35$; $p < 0.001$). There was no effect of *group* ($\chi^2(1) = 0.31$; $p > 0.05$), nor was there a *VOT* by *group* interaction ($\chi^2(1) = 3.04$; p > 0.05). In panel (a) of Figure 5 one can observe two nearly overlapping sigmoid functions, suggesting the two groups identified the resynthesized continuum in a similar manner. With regard to the crossover boundary data, there was no effect of *group* ($\beta = 0.24$, CI low = -0.71, CI high = 1.20, SE = 0.46, $t = 0.54$, p = 0.60). The vertical bars in panel (a) of Figure 5 show that the perceptual boundary for both groups nearly overlap. The model fit to the contrast coefficient slope data did not yield an effect of *group* ($\beta = 0.84$, CI low = -0.04, CI high = 1.72, SE = 0.42, $t = 2.02$, p = 0.06), though the effect approached significance. Panel (b) of Figure 5 plots the CCS slopes in the logistic space. The steepness of the lines suggests both groups had "crisp" boundaries, though it can be observed that the native control group has a slightly steeper slope.

++ INSERT FIGURE 5 ABOUT HERE ++

Interim discussion

The perception experiment was concerned with uncovering how late, sequential language learners develop L2 perceptual strategies. Specifically, the perceptual identification task was designed with the purpose of analyzing how the /b/-/p/ stop contrast was perceived as exposure to the target language increased, and how self-report measures of input and use affect stop perception. The results of the task revealed that the learners did indeed shift their perceptual boundaries with increased exposure to Spanish, though there was no evidence that this shift was modulated by language input nor language use. The analyses did suggest that after 14 days of exposure, learners began to identify the resynthesized stimuli differently from how they had identified the same stimuli two weeks prior. This finding supports the notion that the learners may have begun the process of developing an L2-specific perceptual system. In this case, the contrast in question, /b/-/p/, was one that already existed in their L1. Thus the learning that took place involved the resetting of the perceptual boundary between the segments in this contrast. The analyses found that by the end of the 7-week immersion program—approximately 47 days—, the perceptual boundary of the learners was within the range of the control group of simultaneous bilinguals. Moreover, the learners' linear slope corresponding with the sigmoid functions of the boundary between the two segments was also within the native bilingual range, though it did appear to be slightly less "crisp". Thus far the results of production and perception experiments support the notion that phonetic category development may occur in a relatively short amount of time, at least when L2 use is high and L1 use in minimized.

Production/perception interface in L2 learning

The final study examined the relationship between production and perception in adult L2 learning. Specifically, the present analyses aimed to (1) determine if there was a correlation between production gains and perceptual boundary shifts in late learners of Spanish, and (2) determine if phonetic category development was perceptually driven, or if production gains occurred before perception improved. To shed light on these issues, the longitudinal production and perception data presented in the previous tasks were analyzed together.

Method

Statistical analysis.

Phoneme boundaries.

Phoneme boundaries were calculated for each modality (production, perception). For the production data, the boundary was the mean standardized value for all bilabials (/b/, /p/) produced by a given participant for each session. That is, VOT was normalized as a function of voicing and a mean was then calculated for each individual in each session. Each participant provided one production boundary value per session. The perceptual phoneme boundaries were the 50% crossover values analyzed in the perception task.

Boundary trajectories: motivating GAMMs.

The perception/production boundary data were analyzed using Generalized Additive Mixed Models (GAMM, Sóskuthy, 2017; Winter & Wieling, 2016; Wood, 2006). GAMMs represent an extension to the linear model framework that allow non-linear functions called factor smooths to be applied to predictors. In this sense, the predictors can be classified into two types: parametric terms (equivalent to fixed effects in hierarchical model terminology) and smooth terms. Random smooths are conceptually similar to random slopes and intercepts in the mixed-effects regression framework (Winter & Wieling, 2016). Thus, they allow the by-subject trajectory shapes to vary as a function of a parametric effect and are essential in avoiding anti-conservative models.

Establishing production boundaries.

As we have seen, the calculation of the perceptual boundary is straightforward. This is not true for a production boundary given the fine-phonetic variability found in voicing realizations. Previous work has utilized only voiceless stops as means to make comparisons with perception. In the present analysis, production boundaries were calculated by standardizing the VOT values from both stop categories and averaging them together. To justify this calculation it is necessary to show that the rate of change over time was similar for both segments. Recall from the production task that the learners' voiceless stops decreased by approximately 10 ms by the end of the program and fell within the range of the native bilinguals. For the voiced stops, VOT values decreased by approximately 40 ms by the end of the program and did not fall within the native range. At first glance it appears that the voiced stops showed a larger change over time. However, voiced stops have a wider range of possible values with lead-VOT and short-lag VOT realizations, thus mean change may mischaracterize overall production gains. For this reason VOT values for each segment were standardized in order to put them on the same scale, and subsequently analyzed using a GAMM to compare the change in VOT over time, that is, to analyze the learning trajectory of each segment.

To this end, standardized VOT values were modeled as a function of the parametric term *voicing* (voiced, voiceless) and a non-linear function of *exposure time*. *Voicing* was set as an ordered variable with voiceless stops coded as 0. Cubic regression splines with 7 basis knots were applied (1) as a reference smooth to *exposure time*, (2) as a difference smooth to *exposure time* conditioned on *voicing*, and (3) as a random smooth for each participant conditioned on *exposure time*. This specification uses the voiceless stop trajectory as the baseline and compares it to the voiceless trajectory. Given that the VOT values were standardized, we do not expect a *voicing* effect on the intercept (both levels of *voicing* have an overall mean of 0). The model will, however, be informative regarding the shape of the voiced and voiceless stop trajectories and how they differ from each other in terms of curvature.

The model found no effect for the parametric voicing term, nor the corresponding smoothing terms (see Table A1 for the model output), indicating that the trajectories for voiced and voiceless stops did not differ from each other. Thus the standardized units were averaged together to create the production boundary values which were subsequently combined with the perception boundaries. In a separate GAMM this combination of category boundary data was modeled as a function of modality (production, perception).

Production/perception trajectories.

The model specification for the category boundary data was the same as the previous voicing model and is described again here for the sake of completeness. VOT category boundary values were modeled as a function of the parametric term *modality* (production, perception) and a non-linear function of *exposure time*. *Modality* was set as an ordered variable with perception coded as 0. Cubic regression splines with 7 basis knots were applied (1) as a reference smooth to *exposure time*, (2) as a difference smooth to *exposure time* conditioned on *modality*, and (3) as a random smooth for each participant conditioned on *exposure time*. This specification uses the perception trajectory as the baseline and compares it to the production trajectory. Given that the boundary values were standardized, we again do not anticipate a *modality* effect on the intercept (both modalities have an overall mean of 0), though the model will be informative regarding how the shape of the production and perception learning trajectories differ from each other as a function of *exposure time*.

All analyses were conducted in R using the `mgcv` package (Wood, 2004) for model fitting and `itsadug` for visualization (van Rij et al., 2017). Autocorrelation was inspected visually using autocorrelation function (ACF) plots of model residuals. Significance testing for effects on *modality* were conducted using a combination of t-tests and approximate F-tests on parametric and smooth terms, respectively, in conjunction with nested model comparisons.

Results

Panel (a) of Figure 6 provides a scatterplot of the category boundary data. The vertical axis plots the production boundaries and the horizontal axis plots the perception boundaries. Lower VOT values take darker colors and higher VOT values are mapped to lighter colors. Additionally, the two modalities are mapped to different parts of each individual point. Production values are represented by the outer color of the points, while perception values are represented by the inner color of the points. Exposure time is mapped to each point based on geometric size so that the smallest points imply 0 days of exposure and size increases in parallel with exposure time. One can observe that smaller, lighter points are aggregated in the upper right quadrant of the plot and increase in size and darkness as one moves towards the lower left quadrant. In other words, VOT values for production (vertical axis) and perception (horizontal axis) appear to decrease as exposure increases (point size). Furthermore, the relationship between category boundaries is captured by the regression line plotted in black. One can observe that production and perception boundaries decrease in tandem. The plot, while qualitative in nature, suggests there is a relationship between the two modalities for these learners.

++ INSERT FIGURE 6 ABOUT HERE ++

The modality GAMM explained 48.15% of the variance and 51.48% of the deviance. Nested model comparisons suggested that the parametric and smooth terms on modality significantly improved fit (DF = 3, Difference = 7.11, EDF = 9, $p < 0.003$). The perception boundary varied as a function of exposure time (Reference smooth: EDF = 2.09, Ref. DF = 2.54, F = 8.41, $p < 0.001$). The value higher than 1 for the effective degrees of freedom (EDF) indicates that the trajectory was non-linear. Crucially, the production trajectory differed from the perception trajectory (Difference smooth: EDF = 4.30, Ref. DF = 5.05, F = 4.68, $p < 0.001$). The EDF value indicates that the difference between the trajectories was also non-linear. Panel (b) of Figure 6 plots the modality trajectories (left side) and the estimated differences between them over the time course (right side). The plots corroborate the findings derived from the GAMM. Specifically, one can observe the non-linear time course of both modalities. The perception boundaries shift earlier and at a consistent rate before the slope flattens out around 25 days of exposure. The production boundaries, on the other hand, follow a sigmoid-like trajectory, flattening out around around the same time point as the production trajectory before rising and falling again around the final testing session. The estimated difference plot indicates two exposure time windows of significant differences: from days 2.65 - 13.83 and days 23.36 - 29.98.[5] These are time windows in which the difference in standardized VOT between voiced and voiceless stops is significant at the 0.05 alpha level and are highlighted in the plot by the vertical, discontinuous red lines. The full model summary is available in Table 7.

++ INSERT TABLE 7 ABOUT HERE ++

Interim discussion

The final study presented an analysis of longitudinal data from the production and perception tasks. These data were utilized to calculate category boundaries—for production and perception—for each individual, for each experimental session. The purpose of the present work was to determine if production and perception were related in beginning L2 learners, and to determine if these learners showed perceptual learning before increases in production accuracy (or *vice versa*). The analyses suggest a clear relationship between production and perception in the learners' phonetic behavior. Specifically, phonetic boundaries for both modalities decrease (i.e., shift towards native bilinguals' Spanish boundaries) as exposure to and use of Spanish increases.

Of theoretical importance is that fact that the boundary shifts in speech perception preceded those of speech production. The perceptual boundary shifts occurred early and crossover points decreased at a steady rate over time before flattening out near the end of the program. Importantly, by this point in their development, the L2 learners had

[5] The estimated difference plot extrapolates the numeric time predictor over 100 values ranging from 1 to 42.

perceptual boundaries that fell within the range of the bilingual controls. The significant difference between the modality trajectories supports the notion that production boundary shifts occurred later in time. Thus, the data suggest that production and perception are intimately related in the beginning stages of L2 phonetic category development and that this development is perceptually driven.

General discussion

Summary of findings

The present work examined production and perception tasks, along with an analysis that compared the two modalities. The production task showed that adult L2 learners of Spanish reduced VOT in their production of bilabial stops at the end of the immersion program. Moreover, production gains were modulated by self-reported English use. Specifically, higher English use was associated with less target-like stop production. Overall, the learners reduced aspiration for Spanish /p/ such that VOT was within the range of the bilingual control groups' productions, though they were still slightly higher. The learners also incorporated prevoicing into their production of Spanish /b/, though their realization of the voiced segment was unstable. The results of the production task show that with limited L1 use and high amounts of L2 input learners' pronunciation of Spanish bilabial stops improved after limited exposure time (7 weeks).

The second task examined the learners' perception of bilabial stops in a 2AFC identification task. The analyses showed that learners identified the same VOT continuum differently over the course of the immersion program. Specifically, they were more likely to categorize the resynthesized stimuli as being voiceless as exposure to and use of Spanish increased. The increased voiceless responses led to a phonetic boundary shift to the left, i.e., to a lower crossover point. The boundary shift was consistent with a more native-like perception of the stimuli, that is, more towards the boundary of the bilingual group.

Perceptual categorization was not associated with input or use variables. The results suggest that perceptual learning had taken place. Specifically, the evidence is consistent with the notion that the learners were in the beginning stages of developing language-specific speech perception.

The final analysis directly compared the longitudinal data from the production and perception tasks. The purpose of the comparison was to determine if the two modalities were related in the learner data, and also to uncover if perceptual learning had occurred before the observed production gains (or *vice versa*). The two modalities showed a decrease, or shift, in the phonetic category boundaries as exposure to and use of Spanish increased. Moreover, the analyses showed that the perceptual boundary shifts occurred prior to the production boundary shifts. Thus, the analysis provided evidence supporting the hypotheses that (1) production and perception are indeed related in the beginning stages of L2 learning and (2) that phonetic learning is perceptually driven during this early period of development.

L2 phonetic category development

The results of the production task parallel those of Zampini (1998), who found that native English intermediate/advanced late learners studying Spanish in a public university reduced aspiration in voiceless stops, but still aspirated more than a control group. It may be the case that the learners in question would have become more target-like over a more extended period of time, or that the differences were due to the learners maintaining separate categories for English and Spanish. In either case, the results from the present production task extend Zampini's findings to a domestic immersion context where similar changes occurred in a shorter time span (7 weeks). The present study differs from Zampini (1998) in the voiced segment, /b/, which showed evidence of prevoicing by the end of the program. Specifically, the learners showed a large relative change in VOT, but average values still fell outside the bilingual range, suggesting there was still room for improvement.

However, analyzing the subset of prevoiced stops showed that all participants produced /b/ with lead-VOT some of the time, and, when they did, VOT values were closer to the bilingual range. The advanced learners in Zampini (1998) did not incorporate prevoicing into their productions of /b, d, g/.

The findings presented in the perception task also corroborate those of the perceptual experiments conducted in Zampini (1998). Zampini (1998) showed that the late learners improved their perception of Spanish bilabials over the course of a semester in a traditional classroom setting. Specifically, Zampini (1998) found that perception of English and Spanish stops became more target-like in conjunction with the learners' progress in Spanish. The results from the perception task presented here suggested that the process of perceptual learning may be sped up with increased L2 exposure and minimal L1 use. Concretely, these results showed that 14 days of L2 input and high L2 usage were sufficient for the adult learners of Spanish to shift their perceptual boundaries in a manner consistent with more Spanish-like perception of bilabial stops.

The results also draw parallels with those of Stevens (2001). His pre/post test analysis of Spanish stop production in a foreign immersion context found that adult learners reduced aspiration in voiceless stops, and that this reduction was not observed after one semester in traditional classroom learners. The present study extends these findings to the domestic immersion context, at least for /p/. Stevens (2001) also found that length of stay was positively correlated with production accuracy. Students that spent an entire semester abroad produced more target-like stops. Stevens (2001) posited that this may be due to increased use of Spanish. The present study found that, generally, production became more target-like after around three weeks of exposure, and continued to do so throughout the program. Due to the pre/post test nature of the experimental design in Stevens (2001) it is impossible to say when exactly production gains began to occur. Be that as it may, considered in conjunction with the present work, these findings support the notion that maximizing L2 use and L2 input can accelerate the acquisition of L2 phonology. This affirmation is consistent with similar research that finds that the largest phonetic gains occur during initial stages of learning (Williams, 1979; Munro et al., 2012).

Taken together, the results from the production and perception tasks suggest L2 phonetic learning can take place in an immersion context in a relatively short period of time. The findings presented herein point to phonetic category development for the stop voicing contrasts of Spanish, though this affirmation cannot be confirmed in the absence of data from the learners' L1. Taking the most conservative view, these findings are at the very least consistent with the early stages of phonetic category development.

The present findings also corroborate the notion that the ability to learn novel sounds is maintained throughout the lifespan (Flege, 1995). With regard to the SLM, the fact that the stops of Spanish are phonetically similar to those of English did not appear to impede the learner group from detecting the fine-phonetic detail between them. Contrary to an SLM account, the L2LP model predicts similar contrasts to be easier to learn. The case presented here may be in agreement with the predictions of this model, though the notion of "ease" is opaque, particularly in light of the fact that voiced bilabial production was unstable. Another question to consider is whether or not adult learners are capable of becoming native-like in the perception of their L2. In the context of the present study, the term "native-like" is in reference to the behavior of the simultaneous (native) bilinguals. The findings of the perception experiment revealed that the adult learners' boundaries shifted to within the range of the native bilingual control group. It is important to note, nonetheless, that both the SLM and the L2LP models conceive of the term "native-like" in reference to the production/perception behavior of monolinguals. The SLM predicts native-like attainment will be less likely with increased age due to cross-linguistic interferences which result from the L2 categories residing in the same phonetic space as the L1 categories. The L2LP model, for its part, maintains that native-like perception is indeed possible due to the fact that the L1 sound system is copied and serves as the starting point for L2 perception. With exposure, the boundaries between L2 categories are reset through the GLA, which then operates independent of the L1 sound system. The results from the perception experiment do not support any single model over the others because they do not provide insight regarding the status of the L1 categories after the program.

In sum, it remains to be seen how the perceptual development displayed here would continue over time. Does a newly acquired sound system disappear with diminished L2 use/input? What are the consequences for L1 sound categories? The L2LP predicts that the L1 system will not be influenced by the development of the L2 system. L1-L2 interactions are expected to occur under an SLM account; however, if the L1 system is left intact and operates independently of the newly acquired L2 system, as posited by the L2LP, then the phonemic boundary for English stops should not change. Future research could build upon the findings of the present work by including measures of English production/perception in unambiguous English-sessions at the start and conclusion of the immersion program. This would allow for a clearer understanding of cross-language interference during the beginning stages of learning and over time. Furthermore, L1 category changes would provide evidence that the L1 and L2 perceptual systems might not be separate, as suggested by the L2LP model.

Production/perception interface in the acquisition of L2 phonology

All major theories of speech perception posit a production/perception relationship. The present findings corroborate a long line of research demonstrating a relationship between speech production and speech perception in L1 and L2 acquisition (Smith, 1973; Edwards, 1974; Ingram, 1977; Menyuk, 1977; Flege & Eefting, 1988; Williams, 1979; Flege et al., 1999). Moreover, the results presented here contribute to previous research regarding the production/perception interface by demonstrating the relationship between the two modalities in beginning L2 learners with longitudinal data.

Another question addressed in production/perception analyses dealt with the causal relationship between production and perception in L2 learning. Many researchers believe that perception precedes production in language acquisition, though the literature has shown that this assertion is not without controversy. The analyses presented here showed that the perceptual boundary shifts occurred prior to the changes in production. This is taken as evidence supporting the claim that changes in perception precede accurate production in the beginning stages of L2 acquisition in adults. This finding is in line with many studies suggesting that perception drives production in L1 and L2 acquisition (Williams, 1979; Borden et al., 1983; Neufeld, 1988; Barry, 1989; Grasseger, 1991; Flege, 1993; Rochet, 1995; Llisterri, 1995; Flege et al., 1997; Leather, 1999). Furthermore, perceptually driven L2 learning coincides with the contention that phonetic segments are "[...] produced only as accurately as they are perceived" (Flege, 2003, p. 25). It appears, at least in these data, that perceptual readjustments lead to, or at least precede, changes in speech production.

Regarding evidence suggesting production precedes perception in L2 learning, it remains possible that methodological concerns (see Escudero, 2006) can account for those cases. Alternatively, these findings may simply be the result of a time confound that arises based on the current state of the learners grammar at the moment in which the data are collected. Another possibility is that subtle gains in one modality incite gains in the other in a time-lagged relationship. This further underscores the importance of studying phonological acquisition during the initial stages of learning. Future research could address these matters by taking temporal resolution into account. For instance, one could implement a longitudinal design that collects data in more experimental sessions over a shorter period of time. It is also important to note that the data examined in the present work dealt with the development of bilabial stops. Further support for a perception-first developmental path would be provided by demonstrating that perceptual boundary shifts precede production accuracy in other segments as well. The stop segments under analysis likely represent an "easy" boundary adjustment, given that the phonological contrast is already part of the participants' native phonology. Future investigations ought to extend the aforementioned findings to other L2 segments, such as laterals, nasals, and vowels. This would further our understanding of how the production/perception of different phonetic segments develop over time. Of particular interest is the longitudinal development of spirantization of voiced stops and the acquisition of the rhotic trill, /r/. The former is contrastive in American English (as opposed to allophonic in Spanish), and the latter is not part of the American English phonemic inventory. Thus, both segments provide clear instances where novel category formation would be necessary for production/perception, and neither segment is likely to be explained via boundary resetting. In essence, the acquisition of these segments could provide an interesting point of comparison to stop voicing, and the amount of exposure necessary for category formation to occur could shed light on the relative difficulty of boundary resetting versus learning a new allophonic distribution versus learning a new sound altogether.

Finally, the production/perception analysis showed that the trajectories for each modality differed during two points in the time course (see Figure 6b, right panel), once during the initial point of exposure, and later again between days 23 and 30. This second window merits further consideration. Specifically, at this time, the perceptual boundary trajectory was beginning to flatten out as the production boundary continued to decrease. Interestingly, the trajectories approached one another again. A priori one would not expect the trajectories to stop being significantly different. One would expect the production boundary to continue to decline as the learners incorporated more prevoicing into their production. This pattern is not observed. Conversely, the production trajectory begins to rise around day 30, reaches a peak around day 35, and then continues to fall. This variability may be explained by the participants extra-curricular activities during this time frame of the immersion program. Specifically, the Language Schools were celebrating their centennial anniversary and the language pledge was suspended during one evening for a banquet and party. In other words, the participants spent the night prior to the penultimate testing session (day 35) hearing and possibly speaking English. A posteriori the participants were asked if they had spoken English during the centennial celebration. Most reported that they had and that it had been a cathartic experience. This suggests that the variability in the phonetic behavior of the learners may be explained by L1 language use. This hypothesis is supported by the finding that self-reported use of English was correlated with overall change in VOT in the production data. Importantly, English use did not modulate perceptual categorization. This poses the possibility that speech production is more affected by native language activation than speech perception. Taken together, the results suggest that phonetic category development during the early stages of learning appears to be particularly fragile and susceptible to cross-linguistic interference. That said, the trajectory analysis suggests both voiced and voiceless stops would have continued to improve, though it is likely that production gains would eventually slow, possibly to maintain distinct categories for each language. A likely scenario is that Spanish stop production would become more stable with increased exposure to and use of the L2. Future research should directly compare the domestic immersion learning context with study abroad and the traditional classroom with longitudinal designs. This would ultimately help tease apart the effects of L1/L2 use and the relative importance of different types of target language input, which, in the present work, were operationalized to include various factors, such as L2 feedback, motivation, and attention.

Conclusion

The present investigation analyzed early second language learning in adults. The studies undertaken for this work regarding the ongoing development of the fine-phonetic detail of Spanish stop voicing add to our understanding of the acquisition of L2 phonology in adult learners. The longitudinal data suggest that L2 phonetic category formation can occur abruptly at an early stage of development, and is perceptually driven. Moreover, early, developing L2 sound representations are fragile, and especially susceptible to cross-linguistic interference during the initial stages of learning.

References

Barry, W. J. (1989). Perception and production of English vowels by German learners: instrumental-phonetic support in language teaching. *Phonetica*, *46*(4), 155–168.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Birdsong, D., Gertken, L. M., & Amengual, M. (2012). *Bilingual language profile: an easy-to-use instrument to assess bilingualism.* Retrieved from https://sites.la.utexas.edu/bilingual/

Boersma, P., & Weenink, D. (2018). *Praat: doing phonetics by computer [computer program].* Retrieved from http://www.praat.org/

Borden, G., Gerber, A., & Milsark, G. (1983). Production and perception of the /r/-/l/ contrast in Korean adults learning English. *Language Learning*, *33*(4), 499–526.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception & psychophysics*, *61*(5), 977–985.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, *101*(4), 2299–2310.

Caramazza, A., Yeni-Komshian, G., Zurif, E. B., & Carbone, E. (1973). The acquisition of a new phonological contrast: The case of stop consonants in French-English bilinguals. *The Journal of the Acoustical Society of America*, *54*(2), 421–428.

Casillas, J. V. (2018). lingstuff: Tools and gems for linguistics related research [Computer software manual]. Retrieved from https://github.com/jvcasill/lingStuff (R package version 0.1.1)

Davidson, L. (2018). Phonation and laryngeal specification in American English voiceless obstruents. *Journal of the International Phonetic Association*, *48*(3), 331–356.

Díaz-Campos, M. (2006). The effect of style in second language phonology: An analysis of segmental acquisition in study abroad and regular-classroom students. In *Selected proceedings of the 7th conference on the acquisition of spanish and portuguese as first and second languages* (pp. 26–39).

Edwards, M. L. (1974). Perception and production in child phonology: The testing of four hypotheses. *Journal of Child language*, *1*(2), 205–219.

Elliott, A. R. (1997). On the teaching and acquisition of pronunciation within a communicative approach. *Hispania*, 95–108.

Escudero, P. (2005). *Linguistic perception and second language acquisition* (Unpublished doctoral dissertation). Utrecht University, Utrecht, Holland.

Escudero, P. (2006). Second-language phonology: the role of perception. In M. C. Pennington (Ed.), *Phonology in context* (pp. 109–134).

Escudero, P. (2009). Linguistic perception of 'similar' L2 sounds. In P. Boersma & S. Hamann (Eds.), *Phonology in perception* (pp. 151–190). Berlin: Mouton de Gruyter.

Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition*, *26*(4), 551–585.

Flege, J. E. (1981). The phonological basis of foreign accent: A hypothesis. *Tesol Quarterly*, *15*(4), 443–455.

Flege, J. E. (1987a). A critical period for learning to pronounce foreign languages? *Applied Linguistics*, *8*, 162-177.

Flege, J. E. (1987b). The production of “new” and “similar” phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of phonetics*, *15*(1), 47–65.

Flege, J. E. (1993). Production and perception of a novel, second-language phonetic contrast. *The Journal of the Acoustical Society of America*, *93*(3), 1589–1608.

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. 229–273). Timonium, MD: York Press.

Flege, J. E. (2003). Assessing constraints on second-language segmental production and perception. In *Phonetics and phonology in language comprehension and production differences and similarities* (pp. 319–355). Berlin: De Gruyter Mouton.

Flege, J. E. (2012). The role of input in second language (L2) speech learning. Lodz, Poland.

Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers’ production and perception of English vowels. *Journal of Phonetics*, *25*(4), 437–470.

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), 729–740.

Flege, J. E., & Liu, S. (2001). The effect of experience on adults’ acquisition of a second language. *Studies in Second Language Acquisition*, *23*(4), 527–552.

Flege, J. E., MacKay, I. R. A., & Meador, D. (1999). Native Italian speakers’ perception and production of English vowels. *The Journal of the Acoustical Society of America*, *106*(5), 2973–2987.

Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995). Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America*, *97*(5 Pt 1), 3125–3134.

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.

González-Bueno, M. (1997). Voice onset time in the perception of foreign accent by native listeners of Spanish. *International Review of Applied Linguistics in Language Teaching*, *35*(4), 251–262.

González López, V., & Counselman, D. (2013). L2 acquisition and category formation of Spanish voiceless stops by monolingual English novice learners. In J. Cabrelli Amaro, G. Lord, A. de Prada Pérez, & J. E. Aaron (Eds.), (pp. 118–127). Somerville, MA: Cascadilla Proceedings Project.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds “L” and “R”. *Neuropsychologia*, *9*(3), 317–323.

Grasseger, H. (1991). Perception and production of Italian plosives by Italian learners. In *Actes du xii congrès international des sciences phonétiques* (Vol. 5, pp. 290–293).

Ingram, D. (1977). *Phonological disability in children* (Vol. 2). New York: Elsevier.

Kissling, E. M. (2014). Phonetics instruction improves learners’ perception of l2 sounds. *Language Teaching Research*, 1362168814541735.

Leather, J. (1997). Interrelation of perceptual and productive learning in the initial acquisition of second-language tone. In A. James & J. Leather (Eds.), *Second-language speech: structure and process* (pp. 75–102). Berlin, New York: DE GRUYTER MOUTON.

Leather, J. (1999). Second-language research: An introduction. *Language Learning*, *49*(s1), 1–56.

Lisker, L., & Abramson, A. S. (1964). A Cross-language Study of Voicing in Initial Stops: Acoustical Measurements. *Word*, *20.3*, 384–422.

Llisterri, J. (1995). Relationships between speech production and speech perception in a second language. *Proceedings of the 13th International Congress of Phonetic Sciences*, *4*, 92–99.

López, V. G. (2012). Spanish and english word-initial voiceless stop production in code-switched vs. monolingual structures. *Second Language Research*, *28*(2), 243–263.

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

Mathews, J. (1997). The influence of pronunciation training on the perception of second-language contrasts. In J. Leather & A. James (Eds.), *New sounds 97: Proceedings of the third international symposium on the acquisition of second language speech* (p. 223–229). Klagenfurt: University of Klagenfurt.

Menyuk, P. (1977). *Language and maturation.* Cambridge, MA: MIT Press.

Morrison, G. S. (2007). Logistic regression modelling for first- and second-language perception data. In M. J. Sole, P. Prieto, & J. Mascaró (Eds.), *Segmental and prosodic issues in romance phonology* (pp. 219–236). John Benjamins Publishing Company.

Munro, M. J., Derwing, T. M., & Saito, K. (2012). English l2 vowel acquisition over seven years. *PRONUNCIATION AND ASSESSMENT*, 112.

Nagle, C. L. (2017). A Longitudinal Study of Voice Onset Time Development in L2 Spanish Stops. , *54*(5), 13-23.

Nagle, C. L., Moorman, C., Sanz, C., et al. (2016). Disentangling research on study abroad and pronunciation: Methodological and programmatic considerations. In *Handbook of research on study abroad programs and outbound mobility* (pp. 673–695). IGI Global.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods In Ecology And Evolution*, *4*, 133–142.

Neufeld, G. G. (1988). Phonological asymmetry in second-language learning and performance. *Language Learning*, *38*(4), 531–559.

Oyama, S. (1976). A sensitive period for the acquisition of a nonnative phonological system. *Journal of Psycholinguistic Research*, *5*, 261-283.

Pallier, C., Bosch, L., & Sebastián-Gallés, N. (1997). A limit on behavioral plasticity in speech perception. *Cognition*, *64*(3), B9–17.

Peirce, J. W. (2008). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*(10), 1–8.

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rochet, B. L. (1995). Perception and production of second-language speech sounds by adults. *Speech perception and linguistic experience: Issues in cross-language research*, 379–410.

Sakai, M. (2018). Moving towards a bilingual baseline in second language phonetic research. *Journal of Second Language Pronunciation*, *4*(1), 11–45.

Saville-Troike, M. (2005). *Introducing second language acquisition.* Cambridge University Press.

Sebastián-Gallés, N. (2006). Native-language sensitivities: evolution in the first year of life. *Trends in cognitive sciences*, *10*(6), 239–241.

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.

Sebastián-Gallés, N., & Soto-Faraco, S. (1999). Online processing of native and non-native phonemic contrasts in early bilinguals. *Cognition*, *72*(2), 111–123.

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics*, *3*(03), 243–261.

Smith, N. V. (1973). *The acquisition of phonology: A case study*. Cambridge University Press.

Sóskuthy, M. (2017). Generalised additive mixed models for dynamic analysis in linguistics: a practical introduction. *arXiv preprint arXiv:1703.05339*.

Stevens, J. (2001). Study abroad learners' acquisition of the Spanish voiceless stops. *MIFLC Review*, *10*, 137–151.

Strange, W. (1995). Cross-language studies of speech perception: A historical review. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross language research* (pp. 3–45). Baltimore, MD: York Press.

Sundara, M., & Polka, L. (2008). Discrimination of coronal stops by bilingual adults: the timing and nature of language interaction. *Cognition*, *106*(1), 234–258.

Sundara, M., Polka, L., & Baum, S. (2006). Production of coronal stops by simultaneous bilingual adults. *Bilingualism: Language and Cognition*, *9*(01), 97–114.

Sundara, M., Polka, L., & Genesee, F. (2006). Language-experience facilitates discrimination of /d/-/ð/ in monolingual and bilingual acquisition of english. *Cognition*, *100*(2), 369–388.

van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2017). *itsadug: Interpreting time series and autocorrelated data using gamms.* (R package version 2.3)

Wang, X. (2002). Training Mandarin and Cantonese Speakers to Identify English Vowel Contrasts: Long-term Retention and Effects on Production (Unpublished doctoral dissertation). British Colombia.

Williams, L. (1979). The modification of speech perception and production in second-language learning. *Perception & Psychophysics*, *26*(2), 95–104.

Winter, B., & Wieling, M. (2016). How to analyze linguistic change using mixed models, growth curve analysis and generalized additive modeling. *Journal of Language Evolution*(1), 7-18.

Wood, S. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, *99*(467), 673-686.

Wood, S. (2006). *Generalized additive models: an introduction with r* (2nd ed.). Boca Raton: CRC Press.

Zampini, M. L. (1998). The Relationship between the Production and Perception of L2 Spanish Stops. *Texas Papers in Foreign Language Education*, *3*(3), 85–100.

Zampini, M. L., & Green, K. P. (2001). The voicing contrast in English and Spanish: The relationship between perception and production. In *One mind, two languages bilingual language processing* (pp. 23–48). Malden, Mass; Oxford: Blackwell.

		Use		Input type					Estimated ability		
ID	Age	Sp.	En.	NI	NNI	NNI+	NNI−	NNI=	Speaking	Listening	Overall
101	21	98.3	3.3	45.0	55.0	43.3	56.7	50.0	51	74	60
102	26	93.3	6.7	53.3	70.0	68.3	18.3	65.0	69	69	70
103	28	100.0	16.7	51.7	48.3	78.3	10.0	11.7	44	64	50
104	18	95.0	0.0	48.3	53.3	68.3	30.0	45.0	57	71	59
105	20	70.0	31.7	35.0	70.0	61.7	30.0	66.7	44	57	49
106	34	86.7	13.3	33.3	61.7	58.3	15.0	21.7	50	64	60
107	22	80.0	15.0	33.3	63.3	63.3	13.3	43.3	34	50	40
108	29	93.3	6.7	26.7	71.7	98.3	1.7	0.0	14	19	14
109	19	75.0	23.3	20.0	53.3	43.3	28.3	40.0	37	64	41
110	20	91.7	18.3	58.3	91.7	71.7	46.7	83.3	49	64	53
Avg.	23.7	88.3	13.5	40.5	63.8	65.5	25.0	42.7	45	59	49

Table 1

Averaged self-report values (from weekly assessment questionnaire) of Spanish/English use, estimated native input (NI), non-native input (NNI), non-native input from speakers with a higher level (NNI+), non-native input from speakers with a lower level (NNI−), non-native input from speakers with the same level (NNI=), and estimated speaking/listening/overall ability (in Spanish). All measures represent percentages.

	Questionnaires		Perception		Production	
Session	Demographic	Assessment	2AFC (a)	2AFC (b)	Repetition	Picture Naming
Week 0	✓	✓	✓		✓	
Week 1		✓	✓		✓	
Week 2		✓	✓		✓	
Week 3		✓	✓		✓	
Week 4		✓	✓		✓	
Week 5		✓	✓		✓	
Week 6		✓	✓		✓	
Week 7		✓	✓	✓	✓	✓

Table 2

Timetable of experimental sessions for the learner group. 2AFC (b) and picture naming are not reported.

Term	β	SE	CI low	CI high	Statistic	Pr(>|t|)	
Intercept	32.06	5.43	21.42	42.71	5.90	0.001	*
Day 7	2.59	3.41	-4.09	9.27	0.76	0.460	
Day 14	-3.50	3.41	-10.18	3.18	-1.03	0.310	
Day 21	-8.86	3.41	-15.55	-2.18	-2.60	0.013	*
Day 28	-12.21	3.41	-18.89	-5.53	-3.58	0.002	*
Day 35	-9.60	3.41	-16.28	-2.92	-2.82	0.008	*
Day 42	-10.75	3.41	-17.43	-4.06	-3.15	0.004	*

Table 3

Model output for VOT of voiceless stops as a function of exposure time. The intercept represents VOT on day 0. Parameter estimates from subsequent sessions represent the change in VOT with regard to the intercept.

ID	Day 1	Day 7	Day 14	Day 21	Day 28	Day 35	Day 42	Avg.
101	0.00	26.67	60.00	100.00	86.67	100.00	100.00	67.62
102	0.00	0.00	0.00	0.00	6.67	26.67	40.00	10.48
103	0.00	0.00	0.00	0.00	6.67	20.00	33.33	8.57
104	0.00	0.00	0.00	0.00	6.67	13.33	33.33	7.62
105	0.00	0.00	0.00	6.67	0.00	20.00	33.33	8.57
106	6.67	0.00	13.33	46.67	93.33	66.67	53.33	40.00
107	0.00	0.00	0.00	0.00	13.30	26.67	33.33	10.47
108	5.88	6.67	0.00	100.00	87.50	93.75	100.00	56.26
109	0.00	0.00	0.00	26.67	20.00	6.67	86.67	20.00
110	0.00	0.00	0.00	20.00	33.33	33.33	26.67	16.19
Avg.	1.25	3.33	7.33	30.00	35.41	40.71	54.00	24.58

Table 4

Proportion of prevoiced /b/ *realizations as a function of exposure time.*

Term	β	SE	CI low	CI high	Statistic	Pr(>|t|)	
Intercept	13.54	9.14	-4.37	31.45	1.48	0.149	
Day 7	-2.44	9.38	-20.81	15.94	-0.26	0.797	
Day 14	-5.24	9.37	-23.62	13.13	-0.56	0.579	
Day 21	-31.51	9.38	-49.89	-13.13	-3.36	0.002	*
Day 28	-34.30	9.38	-52.69	-15.92	-3.66	0.002	*
Day 35	-30.11	9.37	-48.48	-11.73	-3.21	0.003	*
Day 42	-38.49	9.37	-56.86	-20.12	-4.11	0.001	*

Table 5

Model output for VOT of voiced stops as a function of exposure time. The intercept represents VOT on day 0. Parameter estimates from subsequent sessions represent the change in VOT with regard to the intercept.

Term	β	CI low	CI high	SE	DF	Statistic	Pr(>|t|)	
Intercept	0.57	-0.01	1.15	0.30	27.81	1.94	0.064	
Day 7	-0.14	-0.71	0.43	0.29	70.00	-0.49	0.627	
Day 14	-0.66	-1.24	-0.10	0.29	70.00	-2.29	0.026	*
Day 21	-0.64	-1.21	-0.07	0.29	70.00	-2.21	0.040	*
Day 28	-0.80	-1.37	-0.23	0.29	70.00	-2.74	0.009	*
Day 35	-0.47	-1.04	0.10	0.29	70.00	-1.62	0.110	
Day 42	-0.80	-1.37	-0.23	0.29	70.00	-2.74	0.009	*
Day 47	-1.07	-1.64	-0.50	0.29	70.00	-3.68	0.001	*

Table 6

Model output for perceptual boundary crossover point as a function of time. The intercept represents perceptual boundaries on day 0. Parameter estimates from subsequent sessions quantify boundary shifts with regard to the intercept.

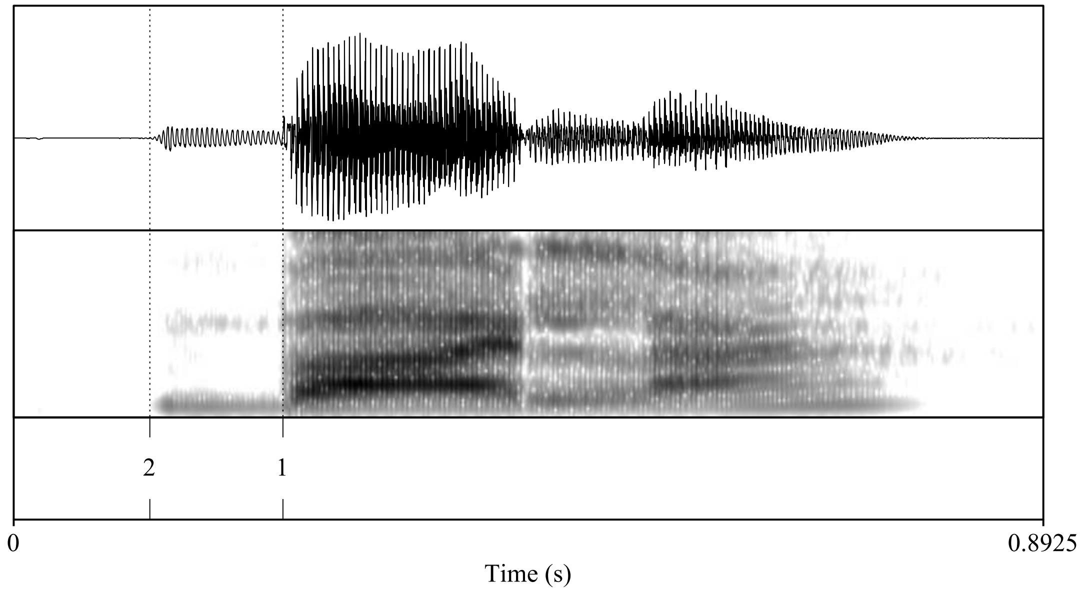
Model output				
Parametric coefficients:				
	Estimate	**Std. Error**	**t value**	**p-value**
Intercept	0	0.08	0	> 0.05
Intercept Δ Production	0	0.098	0	> 0.05
Approximate significance of Smooth terms:				
	EDF	**Ref. DF**	**F**	**p-value**
Reference smooth: exposure time	2.087	2.539	8.408	< 0.001
Difference smooth: production	4.298	5.048	4.684	< 0.001
Random smooth: exposure time x participant	6.016	68	0.14	> 0.05

$R^2 = 0.48$; Deviance explained: 51.48%

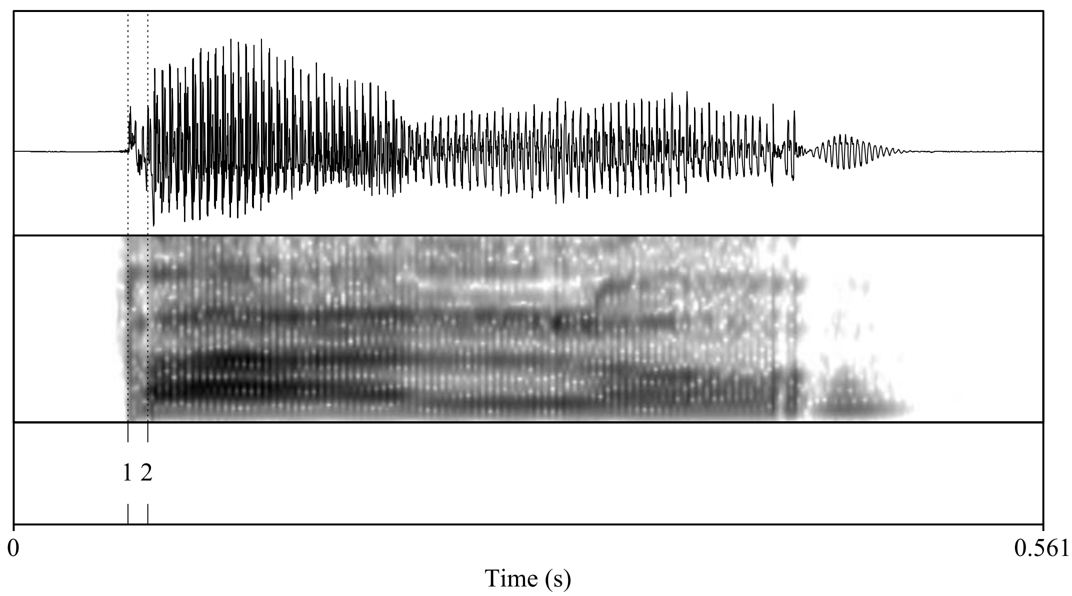
n = 140

Table 7

Summary of production/perception GAMM model output. The full model fit VOT category boundaries as a function of modality.



(a) Spanish *bala* [ˈba.la] (Eng. ‘bullet’).



(b) Spanish *palo* [ˈpa.lo] (Eng. ‘stick’).

Figure 1. Segmenting procedures for Spanish stops. Panels (a) and (b) illustrate pre-voiced and short lag VOT, respectively. The release of the stop is labeled (1) and the onset of modal voicing is labeled (2). VOT was calculated as the time (in ms) from (1) to (2).

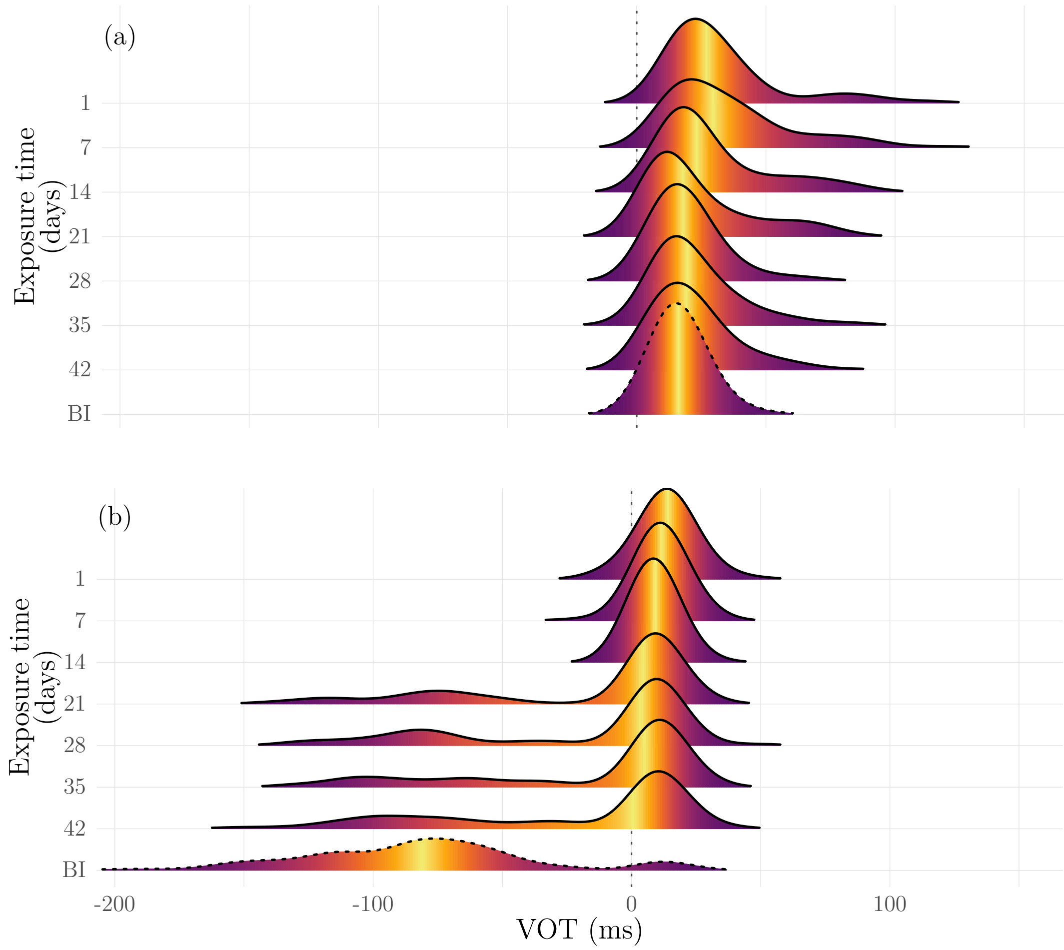


Figure 2. Ridgeline density plots of voiceless (panel ‘a’) and voiced (panel ‘b’) stop VOT (ms) as a function of exposure time. The final ridgeline of each panel, outlined with a discontinuous line, represents the VOT data from the bilingual control group. Lighter colors indicate proximity to the geometric center of gravity of the distribution for each experimental session.

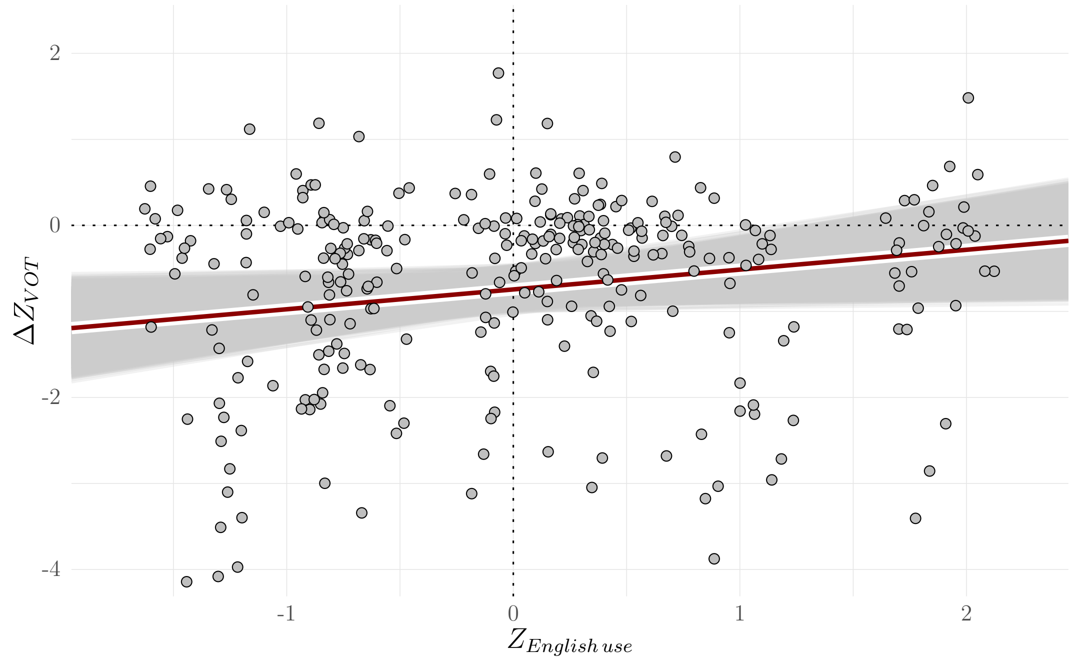


Figure 3. Scatterplot and line of best fit ± 95% CI for ΔZ_{VOT} as a function of *Z-English use*. ΔZ_{VOT} represents the change in standardized VOT for bilabial stops at the end of the immersion program and *Z-English use* represents the average of self-reported time spent speaking English in standardized units. The x-axis is jittered horizontally to show overlapping points.

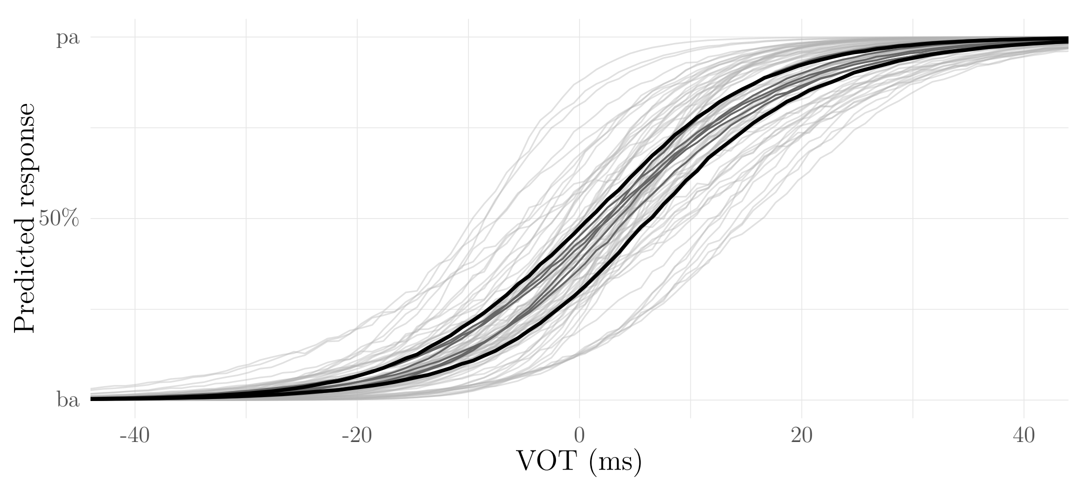


Figure 4. Voiceless responses as a function of VOT. Each line represents an experimental session. The black lines represent the first and last sessions of the program.

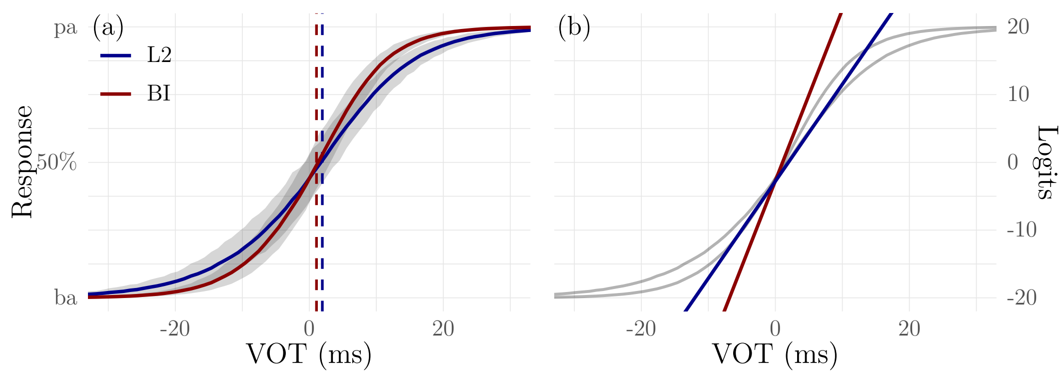


Figure 5. Sigmoidal curves in the probability space (a) and the corresponding contrast coefficient slopes in the logistic space (b) for the learner and bilingual groups. In panel (a), the vertical bars indicate the boundary crossover point for each group. In both panels, the red lines represent the bilingual controls, and the blue lines indicate the learner group.

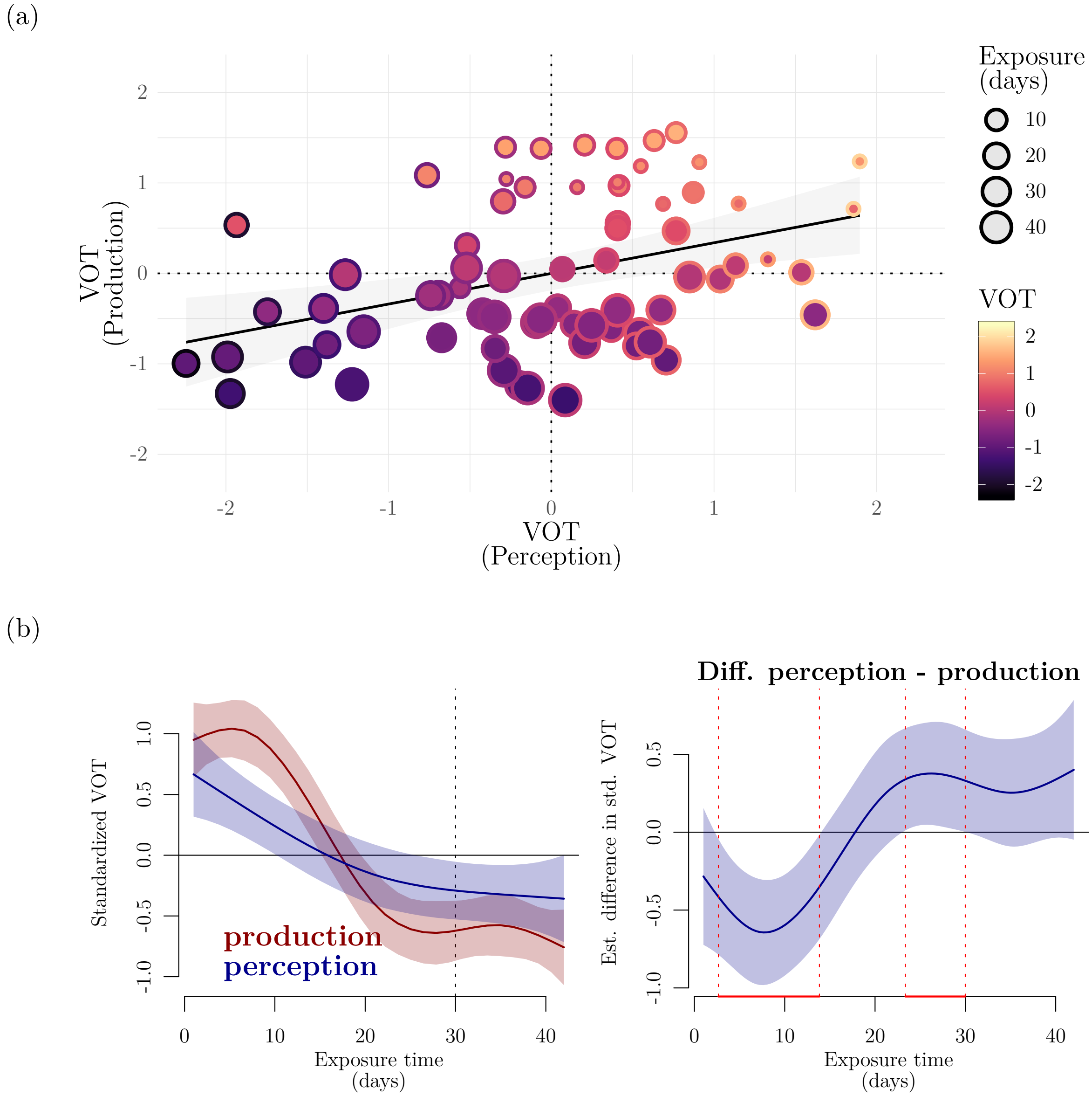


Figure 6. Scatterplot of production and perception boundaries (panel a), non-linear trajectories of production and perception boundaries (panel b, left-side), and estimated difference (voiceless - voiced) in standardized VOT as a function of exposure time (panel b, right-side). In panel (b) the black, discontinuous vertical bar highlights the 30-day mark of exposure time and the red, discontinuous vertical bars represent time windows of significant differences.

Appendix

Model output				
Parametric coefficients:				
	Estimate	**Std. Error**	**t value**	**p-value**
Intercept	0.004	0.075	0.048	> 0.05
Intercept Δ Voiced	0	0.102	0	> 0.05
Approximate significance of Smooth terms:				
	EDF	**Ref. DF**	**F**	**p-value**
Reference smooth: exposure time	1.882	1.95	39.891	< 0.01
Difference smooth: voiced	1	1	2.715	> 0.05
Random smooth: exposure time x participant	3.57	28	0.209	> 0.05

$R^2 = 0.57$; Deviance explained: 59.7%

n = 140

Table A1

Summary of production GAMM model output. The full model fit VOT as a function of voicing.