The role of input in native Spanish Late learners' production and perception of English phonetic segments

James E. Flege, Ratree Wayland

University of Alabama at Birmingham, University of Florida

Abstract. This study evaluated the effect of input variation on the production and perception of English phonetic segments by native Spanish adults who had immigrated to the United States after the age of 16 years. The native Spanish (NS) participants were assigned to three groups of 20 each according to years of English input (years of U.S. residence multiplied by percent English use outside the home). Experiment 1 assessed the perceived relation between English and Spanish vowels. It yielded similar results for the NS groups designated "Low input" (*M* = 0.2 years of input), "Mid" (*M* = 1.2 years) and "High" (*M* = 3.0 years). Experiments 2 – 4 examined English vowel discrimination, vowel production and consonant discrimination. Apart from a modest improvement in vowel discrimination, these experiments showed little improvement as years of English input increased. One possible explanation for the essentially null finding of this study is that input matters little or not at all when an L2 is learned naturalistically following the closure of a critical period. Another possibility is that adequate native speaker input is crucial for L2 speech learning but the input differences evaluated here were insufficient to yield measurable improvements in performance. We conclude the article by illustrating a new technique that might be used to choose between these competing explanations.

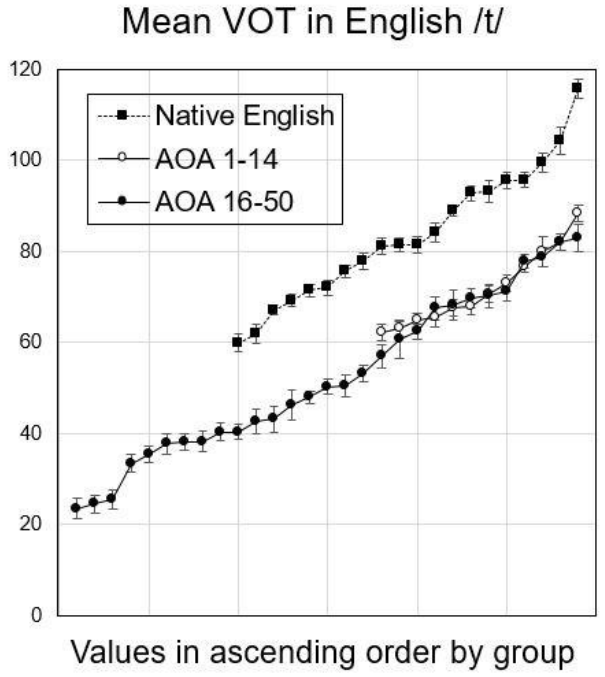
Key words: L2, English, Spanish, production, perception, phonetic segments, SLM, critical period, ESM, EMA

Lenneberg (1967, p. 176) proposed that a critical period exists for L2 speech learning. His proposal derived from the observation that most people who began learning their L2 after puberty speak it with a foreign accent. For Lenneberg, foreign accent in the speech of individuals who might be called "post-Critical Period" learners was an undesired consequence of normal neural maturation. However, the exact neurological underpinnings of the hypothesized critical period remain unclear (Bruer, 2001, p. 19). And while Lenneberg's observation (1967, p. 176) regarding foreign accent was correct, research that has gone beyond a superficial level of observation has raised three issues concerning the basis of the proposed critical period.

First, many adults who began learning an L2 as young children and have used it as their primary language for decades speak their L2 with a foreign accent (e.g., Flege et al. 1997). Second, strength of foreign accents continue to increase long after the hypothesized closure of a critical period (e.g., Flege & MacKay, 2011, Figure 2). Most importantly, Lenneberg's (1967) proposal that normal neural maturation in some way impedes a fundamental human activity, learning speech, ignored the fact that individuals who achieve fluency in an L2 are bilinguals who normally cannot completely prevent their two partially overlapping phonetic subsystems from interacting with one another. Strength of foreign accent in the L2 is inversely related to strength of L2-inspired foreign accent in the native language (Yeni-Komshian et al., 2000).

Foreign accents derive in part from non-nativelike production of the vowels and consonants making up L2 words. That being the case, the critical period hypothesis (CPH) derived from Lenneberg's (1967) observation has been applied to the learning of L2 vowels and consonants. Consider, for example, the wide range of voice onset time (VOT) values obtained in a study examining the influence of lexical factors on the production of phonetic segments in an L2 (Flege et al., 1998). Participants in the lexical study were native Spanish (NS) adults who arrived both before and after the hypothesized closure of a critical period. Each of the mean values shown in Figure 1 are based on productions of 60 /t/-initial English words.

Figure 1. The mean voice onset time (VOT) values drawn from a study by Flege et al. (1998). The means are based on the production of 60 English words by native speakers of English and by native Spanish learners of English differing in age of arrival (AOA) in the United States. The error bars bracket +/- 1 SE. (add 'y' axis label)



All 12 Early learners, but only 11 of the 29 Late learners who participated in the lexical study managed to produce mean VOT values falling within the range of mean values obtained for 20 English monolinguals. Some might interpret the difference in proportion of "native like" Early and Late learners as support for the existence of a critical period. It is reasonable to ask, however, why 38% of the Late learners in this data set were native-like. Lenneberg (1967, p. 176) proposed that following the closure of a critical period L2 learners are unable to make "*automatic*" use of input "*from mere exposure*" (see also Stölten et al., 2014, p. 444). Perhaps *all* of the Late learners in the sample suffered the ill effect of having passed a "critical period", but those with special aptitude for auditorily detecting cross-language differences in VOT overcame the ill effects of having passed a critical period for L2 speech learning.

The results obtained by Flege and Hammond (1982), however, call into question Lenneberg's (1967) input hypothesis. These authors evaluated young adults' awareness of "sub-phonemic" differences between languages. Sensitivity to cross-language VOT differences was evaluated by having native English college students in Florida read English sentences with a feigned Spanish accent. The students who were most familiar with Spanish-accented English shortened VOT in English /t/ to the values often seen in Spanish-accented English. Not only did they detect sub-phonemic VOT differences, they stored this information in long-term memory and were able to use this sensory based information to guide motoric output.

Many researchers acknowledge that Early learners generally receive more and better L2 input than Late learners do (e.g., DeKeyser & Larson-Hall, 2005; Moyer, 2008), and so an input interpretation of the VOT results shown in Figure 1 cannot be ruled out. In addition to explaining differences between Early and Late learners, input differences might also be the key to understanding the inter-subject variability among the Late learners seen in Figure 1.

To better understand the inter-subject variability in Figure 1 we carried out a post-hoc analysis which focused on the 12 Late learners each who produced the longest and shortest VOT values. The subgroups were balanced in terms of gender (3 of 12 participants in each group were male) and overall competence in English (11 members of both groups reported Spanish to be the better of their two languages, and one member indicated equal competence in the two languages). The two subgroups did not differ in terms of self-rated ability to pronounce English (p = .16 by Mann-Whitney *U*-test) or estimates of how difficult it had been to learn English (p = .478).

Our post-hoc analysis revealed that the subgroups of Late learners did not differ in chronological age, age of arrival in the United States or length of U.S. residence (p > .05). However, the Late learners who produced VOT accurately reported a more frequent overall use of English at the time of testing than those who produced English stops with Spanish-like VOT (M = 60% vs 41%, F (1,22) = 4.79, p = .039). It is important to note, however, that neither self-estimated percent English use nor AOA correlated significantly with the VOT values obtained for all 29 Late learners (percent use: r = .32, p = .086; AOA r = -.08 p = .649).

There is controversy as to the relative importance of input for L2 speech learning. Some researchers regard input as being far less important than age of first exposure. For example, DeKeyser and Larson-Hall (2005, p. 88) suggested that "input plays a very limited role" in predicting the outcome of L2 learning once variation in the age of L2 learning has been controlled. DeKeyser (2000, p. 519) claimed that variation in L2 input cannot explain age effects because "it is precisely in the linguistic domain where input varies least – phonology – that age effects are most readily apparent".

According to Flege (2018, see also Flege, 2008), on the other hand, the quantity and quality of input is far more important to successful L2 speech learning than is the age of first exposure to an L2. To illustrate this line of reasoning, consider research examining VOT in the production and perception of English stops by groups of Early learners tested in Texas (Flege, 1991a; Schmidt & Flege 1995, 1996; Flege & Schmidt, 1995; Flege et al., 1996) and in Puerto Rico (Flege & Eefting, 1987, 1988). The Texas Early learners produced and perceived English stops in a native-like fashion whereas the Early learners tested in Puerto Rico produced (and perceived) stops with VOT values that were intermediate to the values obtained from Spanish and English monolinguals. Flege (2018) inferred that the difference between Early learners tested in Texas and Puerto Rico was due to the quality of input that the Early learners had received. Whereas the Early learners in Puerto Rico usually heard English spoken with a Spanish accent by fellow native speakers of Spanish the Early learners tested in Texas usually heard English spoken by native speakers of English.

One reason for a divergence of views regarding the role of input may be an over-reliance by some investigators on length of residence (LOR). This variable has been used to index overall amount of L2 input but is often not suitable for this purpose. LOR is not related linearly to the quantity of L2 input that immigrants receive because not all immigrants begin using L2 immediately (e.g., Flege et al., 1995, Table I) or use the L2 on a regular basis (Moyer, 2008, p. 162). The results of Flege and Liu (2001) suggested that LOR provides a useful index of quantity of L2 input only for those immigrants who have both the opportunity *and* the need to use the L2 on a regular basis.

The data now in hand do not offer a clear understanding of the role of input. Current research generally relies on participants' self-estimates of L2 use that have been obtained via written language background questionnaires. Participants are asked, for example, to estimate their overall L2 use in various time periods (e.g., currently, the last 5 years), in various portions of their waking hours (e.g., at work, at home, socializing with friends) or in specific social contexts (e.g., when shopping, when talking to a neighbor, etc.). Moreover, most previous L2 research has examined groups of participants selected on the basis of an age factor (often, age of arrival in a host country where an L2 must be learned) rather than comparing groups selected on the basis of input.

In summary, the importance of input for L2 speech learning is uncertain. While some researchers (e.g., Flege, 2008) have suggested that an abundance of native speaker input is needed for successful L2 speech learning others have suggested that input plays a minor role, at least in comparison to the maturational state of learners at the time of first exposure to the L2. The aim of this study, therefore, was to provide additional information regarding the effect of variation in quantity of L2 input on the production and perception of English phonetic segments.

The present study

This study examined 60 native Spanish speakers having a mean age of 31 years (*range* = 18-47). Like many young native Spanish (NS) adults who immigrate to the United States, most of our participants had learned English through a "combination of classroom instruction from non-native teachers, followed by naturalistic exposure" (Garcia & Froud, 2018, p. 84).

Of the 60 NS adults recruited for this study, 53 reported having studied English at school in their country of origin before arriving in the U.S. Formal instruction in English for these participants began at an average age of 11.9 years and continued for an average of 6.4 years (*range* = 1-16 years). We will nevertheless refer to all of the NS participants tested here as "Late learners" because the first *regular contact with spoken English* for all 60 NS participants began upon arrival in the U.S. after the age of 16 years.

To evaluate the role of quantity of L2 input we assigned the NS participants to groups of 20 each based on a variable we call "Years of Full-time Input" (or, "input", for short). This variable was calculated for each participant by multiplying length of residence in the United States (LOR, in years) by the immigrant's self-estimated percentage use of English outside the home (specifically, at work and with friends) at the time of test. The resulting three groups, designated the relatively "Low", "Mid" and "High" input groups, were estimated to have received 0.2, 1.2, and 3.0 years of spoken English input, respectively.

The 60 NS Late learners took part in four experimental tasks in a single session lasting about 1.5 hours. One was a perceptual mapping experiment which investigated how the NS participants perceptually related English vowels to the five vowels of Spanish, /i e a o u/. We also examined the NS participants' discrimination of English vowels, their accuracy in producing English vowels, and their discrimination of English consonants. In these last three tasks the NS participants' performance was compared to that of 20 native speakers of American English.

General Methods

Participants

Eighty normal-hearing adults were recruited in Birmingham, Alabama via newspaper ads and personal contact. The 20 native English (NE) speakers (10 male, 10 female) had a mean age of 30 years (*range* = 23-44). Of these, 13 were from the American South (mostly Alabama), four from the Midwest, two from the Northeast and one from the Western portion of the United States.

The NS participants were similar in age to the NE participants (*M* = 31 years, *range* = 17-47). They came from 15 different countries: Mexico-27, Colombia-8, Ecuador-4, Venezuela-4, Costa Rica-3, Peru-3, Uruguay-3, Puerto Rico-2 and one each from Chile, Panama, Spain, Honduras, Argentina and the Dominican Republic. The NS participants arrived in the U.S. between the ages of 17 and 46 years (*M* = 28.1 years) and had been living there from five days to 6.0 years (*M* = 2.6 years) when tested. The NS participants' self-reported English use varied considerably, ranging from 1% to 100% (*M* = 55%).

The variable used for group assignment, "Years of Full-time English Input" (or "input", for short) was calculated for each participant by multiplying length of residence in the United States (LOR, in years) by self-reported percentage use of English outside the home at the time of test. We reasoned that an individual who used English 80% of the time for 3.0 years while living in the U.S. was likely to have had more experience speaking and hearing English spoken than someone who reported using English 10% of the time for 6.0 years. The input variable calculated here for two such participants would be 2.4 vs 0.6 years. Although this represents an improvement over use of LOR, we readily admit that it is less than an ideal estimate of input because: it is based on an unverified estimate of percentage use; it takes into account only current use; and it ignores the potentially important influence of Spanish-accented English. (We return to these issues at the end of the article.)

The NS participants were assigned to three non-overlapping groups of 20 each based on years of input. The mean values for the groups designed "Low", "Mid" and "High input" were 0.2 years, 1.2 years and 3.0 years, respectively.

The participants' overall degree of perceived foreign accent in English sentences was evaluated in a preliminary study using a 9-point rating scale. Not surprisingly, the ratings obtained for all three groups of Late learners (Low *M* = 3.4, *SE*M =.22; Mid *M* = 3.5, *SEM* = .27, High *M* = 4.54, SEM = .41) were significantly lower than those obtained for the NE speakers (*M* = 8.8, *SEM* = .06) [*F* (3, 58) = 38.1, *p* < .0001]. A Tukey post-hoc test indicated, however, that the foreign accents of the Low and Mid groups were significantly stronger than those of members of the High group (*p* < .05).

Procedures

All testing took place in a sound booth located on the campus of the University of Alabama at Birmingham. The participants began by responding to a language background questionnaire. Their hearing was then evaluated using a 25-dB HL criterion at 250, 500, 1,000, 2,000 and 4,000 Hz frequencies in both ears. Two participants who failed to pass the hearing screening were replaced by two others from the same population. The participants then took part in the following experimental tasks: (1) Production of English words, used for the analysis of vowel production accuracy; (2)

Repetition of English sentences, later used to evaluate overall degree of foreign accent; (3) Discrimination of English vowels; (4) Discrimination of English consonants; (5) Cross-language mapping of Spanish and English vowels.

The cross-language mapping experiment, which evaluated how the NS participants related English vowels to the five vowels of Spanish, was ordered last in the sequence of experimental tasks to ensure that the presentation of both Spanish and English vowels in a single context would not influence the NS participants' production or perception of English vowels (tasks #1 and #3). However, the result of the Cross-language mapping task will be presented first because its results shed light on the production and discrimination of English vowels.

Experiment 1: Cross language mapping

The aim of this experiment was to examine the effect of variation in L2 input on the perceived relationship of the 11 English vowels /i ɪ e^{ɪ} ɛ æ ɑ ʌ u ʊ o^{ʊ} ɝ/ to the five vowels of Spanish, /i e a o u/.

According to the Speech Learning Model, or SLM (Flege, 1995) L2 learners perceptually relate sounds heard in the L2 to pre-existing L1 categories via the mechanism of equivalence classification. Important unanswered questions are whether equivalence classification operates at the time of first exposure to an L2 and whether it applies to all L2 vowels and consonants. The SLM proposes, however, that patterns of inter-lingual may change as a function of input. Specifically, as learners become more aware of phonetic differences between sounds in the L1 and L2 inventories, the influence of equivalence classification may diminish. By hypothesis, new phonetic categories created for L2 sounds are not perceptually linked to pre-existing L1 phonetic categories. The SLM does not specify, however, how *much* L2 input learners need in order to recognize that certain L2 vowels are out-of-inventory (i.e., not realizations of a pre-existing L1 category). Nor does the SLM specify how *distant* in phonetic space the L2 vowels must be for this to occur, or *how long* the formation of new phonetic categories might take.

The approach used here to evaluate patterns of inter-lingual identification was inspired by the pioneering work of Robert Scholes (1967) at the University of Florida (see also Flege, 1991b). Scholes presented isolated synthetic vowels differing in first and second formant frequencies (F1, F2) to native speakers of English and Spanish. The synthetic vowel stimuli that were normally heard by the NE listeners vowels as /i ɪ e^{ɪ} ɛ æ ɑ o^{ʊ} u/ were identified. The NS listeners used Spanish key words to classify all stimuli in the two-dimension grid of synthetic stimuli as one of the five vowels of Spanish or else indicated that a stimulus was not a Spanish vowel.

The vowels heard by the NE listeners as /i/, /ɪ/, /e^{ɪ}/, /ɑ/, and /o^{ʊ}/ (the vowels in words such as *beat, bit, bait, bought, boat,* respectively) were heard 100% of the time by the NS listeners as a Spanish vowel (/i/, /i/, /e/, /a/ and /o/, respectively). The NS listeners used the "not Spanish" label some of the time, however, for vowel stimuli classified by the NE listeners as /æ/, /u/ and /ɛ/. For these stimuli, the "not Spanish" response was used 89%, 40% and 28% of the time, respectively. These results suggest that, of the vowels examined, /æ/ was the English vowel most likely to be recognized by NS learners of English as being a "new" (not Spanish) vowel.

Method

Stimuli

The English and Spanish vowel stimuli used here were produced by five adult male native speakers each of English and Spanish. The native English (NE) speakers were from the Southeastern portion of the United States whereas the native Spanish (NS) speakers came from three different countries (Chile-3, Cuba-1, Spain-1). The NE speakers produced English vowels /i ɪ e$^{\text{ɪ}}$ ɛ æ ɑ ʌ u ʊ o$^{\text{ʊ}}$ ɝ/ in a /b_bo/ context. The NS speakers produced Spanish /i e a o u/ in the same phonetic context.

The disyllables were digitized at 22.05 kHz with 16-bit amplitude resolution. All portions of the signal following closure of the intervocalic /b/ were digitally removed. A single token of pre-voicing from one NS speaker was digitally cross-spliced onto all of the remaining tokens. The resulting 80 edited CVC stimuli (11 English vowels x 5 replicate tokens + 5 Spanish vowels x 5 replicate tokens) were then normalized for peak intensity.

Procedure

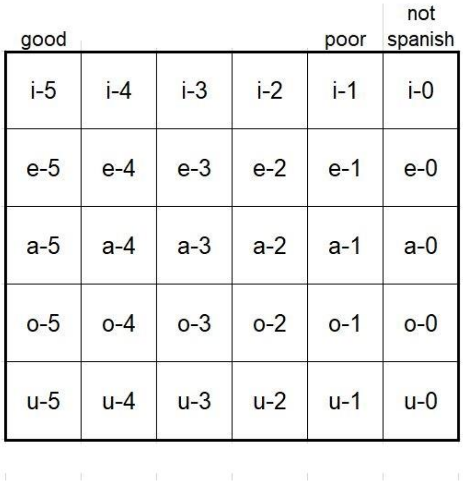
The 80 stimuli were randomly presented one time each via headphones to the 60 NS participants. The participants were told that they would hear both Spanish and English vowels and that their job was to evaluate each vowel in terms of the five vowels of their L1, Spanish.

The procedure normally used to evaluate cross-language perceptual mapping (e.g., Guion et al., 2000) is to present each stimulus twice, the first time for classification in terms of an L1 category and the second time for "goodness of fit" to the L1 phonetic category just selected using a 5-point rating scale. This procedure requires that participants hold their classification judgment in working memory until they hear the second presentation of the same stimulus.

A problem with this procedure is that it will at times force listeners to make counter-intuitive judgments, that is, to rate an L2 vowel stimulus for "goodness of fit" to an L1 vowel category to which the L2 stimulus is *not* perceptually linked. For example, our stimulus set included five productions of the English rhotic (/r/-colored) vowel /ɝ/. These vowel stimuli differed from the other English vowels examined here, and from all five Spanish vowels, in terms of third formant (F3) frequency. We anticipated that our NS participants would not identify the English [ɝ] stimuli in terms of any Spanish vowel. If so, obtaining a goodness of fit judgment for the [ɝ] stimuli would be inappropriate.[1]

[1] Our expectations regarding /ɝ/ derived from research with Italian, a Romance language that has no rhotic (/r/-colored) vowel. Flege & ('&' or 'and') MacKay (2004) examined the cross-language mapping of English vowels to vowels in the Italian inventory by Italian adults having little prior exposure to spoken English. The Italians classified English vowels in terms of one of the seven vowels of Italian (first presentation) and then rated the same vowel stimuli for goodness to the selected category for goodness of fit using a 5-point scale (second presentation). Reponses to English /ɝ/ differed from those obtained from all of the other English vowels, both in terms of the dispersion of response types given and in terms of the very low goodness of fit ratings accorded the /ɝ/ stimuli.

Figure 2. The response grid used in Experiment 1.The rows corresponded to the five vowels of Spanish whereas the selection of differing columns within a row indicated goodness of fit (5 = good, 1 = poor). Responses in the sixth column, “not Spanish”, were used for vowels judged to be out-of-inventory. See text.



good				poor	not spanish
i-5	i-4	i-3	i-2	i-1	i-0
e-5	e-4	e-3	e-2	e-1	e-0
a-5	a-4	a-3	a-2	a-1	a-0
o-5	o-4	o-3	o-2	o-1	o-0
u-5	u-4	u-3	u-2	u-1	u-0

The stimuli described earlier were randomly presented just one time each in the present experiment. The NS participants clicked one of the 30 boxes in the grid shown in Figure 2 to indicate their perceptual judgments. The five rows of the grid corresponded to the Spanish vowels /i/, /e/, /a/, /o/ and /u/ (orthographically <i e a o u>). The columns were used to indicate perceived goodness of fit. Specifically, the values in the first five columns indicated decreasing goodness of fit along a 5-point scale ranging from “good” to “poor”. The NS participants were told to click one of these columns if they heard a Spanish vowel, but to click the sixth column, marked “not Spanish” (0), if they judged a vowel stimulus to be out-of-inventory, that is, not a Spanish vowel.

The participants were required to respond to each stimulus. They could replay a stimulus but not change a response once given. The interval between successive trials was fixed at 1.0 sec. The NS participants were not trained on the task. However, they participated in a short practice session prior to the experiment to familiarize them with use of the response grid (Figure 1). Also, the 80 experimental trials were preceded by five practice items that were duplicated from the end of the randomized list. Responses to the practice trials were not analyzed.

Results

Classification

The 300 responses obtained for each of the five Spanish and 11 English vowels (5 replicate tokens x 3 groups x 20 participants) were tabulated. The cross-language mapping data shown in Table 1 were based on the row in the response grid that was selected, ignoring variation in goodness of fit.

As expected, the modal categorization of the Spanish vowels /i/, /e/, /a/, /o/ and /u/ were the five letters that are consistently used to write the five Spanish vowels, <i>. <e>, <a>, <o> and <u>. There was little variation in how members of the three NS groups classified the Spanish vowel stimuli.

Table 1. Classification of five Spanish (“S”) vowels and 11 English (“E”) vowels in terms of the five vowels of Spanish by three groups of native Spanish speakers differing in English input (L = Low, M = Mid, H = High, see text). The modal Spanish response categories for each group are boldface.

	Spanish response category														
	/i/			/e/			/a/			/o/			/u/		
Stimulus	**L**	**M**	H	L	M	H	L	M	H	L	M	H	L	M	H
S-i	**93**	**97**	**93**	7	3	6	0	0	0	0	0	0	0	0	1
S-e	19	18	12	**80**	**74**	**84**	0	6	3	1	0	1	0	2	0
S-a	0	1	1	1	0	0	**75**	**78**	**79**	19	15	19	5	6	1
S-o	0	1	0	0	0	1	0	0	1	**80**	**85**	**82**	20	14	16
S-u	0	0	1	0	0	2	0	0	0	6	11	7	**94**	**89**	**90**
E-i	**92**	**92**	**94**	8	7	3	0	0	2	0	0	0	0	1	1
E-e$^{\text{ɪ}}$	19	10	8	**77**	**65**	**80**	4	25	12	0	0	0	0	0	0
E-ɑ	0	0	0	0	0	1	**75**	**64**	**72**	18	29	25	7	7	2
E-o$^{\text{ʊ}}$	0	0	0	0	1	0	3	1	6	**83**	**84**	**85**	14	14	9
E-u	0	0	0	0	0	0	0	0	0	3	6	4	**97**	**94**	**96**
E-ɪ	**44**	**62**	**51**	**44**	32	42	1	3	2	1	0	2	10	3	3
E-ɛ	4	0	2	**53**	**70**	**78**	33	24	16	5	1	0	5	5	4
E-æ	0	3	0	3	5	9	**95**	**87**	**90**	1	0	1	1	5	0
E-ʌ	1	1	0	0	1	3	40	35	37	**43**	**51**	**51**	16	12	9
E-ɝ	13	4	3	28	30	**41**	0	4	4	15	14	20	**44**	**48**	32
E-ʊ	0	1	2	2	3	9	4	2	9	19	26	23	**75**	**68**	57

English /i/, /e$^{\text{ɪ}}$/, /ɑ/, /o$^{\text{ʊ}}$/ and /u/ have traditionally been considered to be the “counterparts” of Spanish /i/, /e/, /a/, /o/ and /u/. Consistent with this view, the modal categorizations of the English [i], [e$^{\text{ɪ}}$], [ɑ], [o$^{\text{ʊ}}$] and [u] were Spanish <i>. <e>, <a>, <o> and <u>, respectively. Once again, there was little variation across the three NS groups. Our results are not completely in accord with the traditional account, however. Spanish /a/ is a low central vowel. The English [ɑ] stimuli were classified as Spanish /a/ somewhat less often (*M* = 70%) than were realizations of the low front English vowel /æ/ (*M* = 91%).

There was less consistency in the cross-language mapping results for English /ɪ ɛ ʌ ɝ ʊ /. The

English [ɛ] and [ʊ] stimuli were usually classified as Spanish /e/ and /u/ (67% of instances for both vowels). The English [ɪ] stimuli were classified somewhat more often as Spanish/i/ than /e/ (*M* = 52% vs 39%). This small difference was due to the responses of the Mid and High groups inasmuch as members of the Low group used the two labels equally often (*M* = 44%).

English /ʌ/ and /ɝ/ differ substantially from any Spanish vowels in that both are mid central vowels (in terms of F1 and F2 frequencies) and /ɝ/, a rhotic (r-colored) vowel, differs from all Spanish vowels in terms of its low F3 frequency. Averaged over groups, the [ʌ] stimuli were identified most often as the back mid Spanish vowel /o/ (*M* = 48%) and somewhat less often as the low mid Spanish vowel /a/ (*M* = 37%) or the high back Spanish vowel /u/ (*M* = 12%). The English [ɝ] tokens, perhaps not surprisingly given this vowel's phonetic "strangeness", were classified in terms of *all five* Spanish vowels (/u/ 41%, /e/ 33%, /o/ 16%) /i/ 7% and /a/ 3%).

Not Spanish

We obtained 3,300 responses to the English vowel stimuli from the NS participants (3 groups x 20 participants x 11 vowels x 5 replicate tokens). Of these, 835 (25.3%) were "not Spanish" responses.

Figure 3. The frequency of times that members of the three native Spanish groups judged the English vowel stimuli to be out-of-inventory, that is, "not Spanish" vowels.

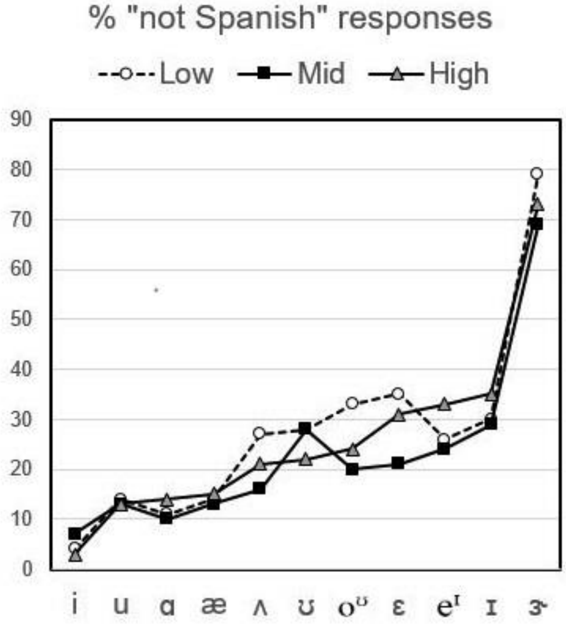


Figure 3 shows the mean percentage of "not Spanish" responses given by members of the three NS groups. The frequency of "not Spanish" responses varied considerably across the 11 English vowels, ranging from 4.7% for /i/ to 73.7% for /ɝ/. There were no systematic differences evident across the three groups. A non-parametric Kruskal-Wallis test indicated that the frequencies observed for the Low, Mid and High groups (301 vs 250 vs 284) did not differ significantly, *H*(2) = 2.03, *p* > .05.

Of the 1,500 responses obtained for the five Spanish vowels, the "not Spanish" response was used

144 times (9.6%). Once again, the difference in frequencies across the three groups (52 vs 36 vs 56) was non-significant, *H* (2) = 5.17, *p* > .05). These "not Spanish" responses to vowels produced by native speakers of Spanish were likely to be due to dialectal differences between the talkers who produced the vowels and the NS participants' own pronunciation of those vowels.

Goodness of fit

The median of five rating given by each NS participants to each of the 11 English and five Spanish vowels was calculated, ignoring how the vowels were classified. The median ratings were then examined in a series of non-parametric tests. In no instance did differences between the three NS groups reach significance at the .05 level.

Figure 4 shows the mean of the median ratings obtained for the five replicate tokens of each English vowel. As expected, the highest ratings were obtained for the English [i] stimuli, the lowest ratings for the English [ɝ] stimuli. The English [e$^{\text{ɪ}}$] and [o$^{\text{ʊ}}$] stimuli were accorded lower ratings than the English [i] stimuli because they are produced with far more tongue movement than their Spanish counterparts. The lower ratings obtained for the English [u] stimuli arose because these stimuli were "fronted", that is, produced with higher F2 frequencies than Spanish /u/.

Figure 4. The goodness of fit ratings obtained for 11 English vowels in Experiment 1. The median of the rating given by each participant to the five replicate (should be 'replicated'?) tokens of each English target vowel was determined. The means shown here are averages of 60 median ratings. The error bars bracket +/- 1 SE.

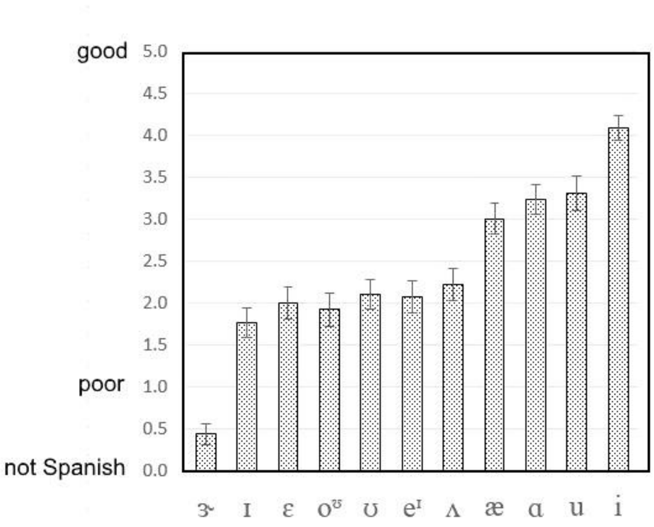


Figure 4 shows the mean of the median ratings obtained for the five replicate tokens of each English vowel. As expected, the highest ratings were obtained for the English [i] stimuli, the lowest ratings for the English [ɝ] stimuli. The English [e$^{\text{ɪ}}$] and [o$^{\text{ʊ}}$] stimuli were accorded lower ratings than the English [i] stimuli because they are produced with far more tongue movement than their Spanish counterparts. The lower ratings obtained for the English [u] stimuli arose because these stimuli were "fronted", that is, produced with higher F2 frequencies than Spanish /u/.

Despite their proximity in vowel space, the NS participants gave substantially lower ratings to the English [ɪ] than [i] stimuli (*M* = 1.8 vs 4.1). To evaluate the role of input, we calculated a difference score between each NS participant's median rating of the five replicate [ɪ] and [i] tokens. The difference scores obtained for the three NS groups did not differ significantly, *H* (2) = 1.96, *p* = .38. Nor did the three groups differ in terms of the difference scores calculated for the English [ɪ] and the

Spanish [i] stimuli, *H* (2) = 1.11, *p* = .57. These findings suggest the cross-language difference between the /ɪ/ of English and the nearest L1 and L2 vowels (in both cases /i/) is readily evident to NS adults from the time they first begin hearing English on a regular basis. It seems that there is no need for a period of "discovery" as Flege (1995) proposed.

In summary, this experiment examined how 11 English vowels were perceptually related to the five vowels of Spanish. There were important differences in the cross-language mapping patterns across the 11 English vowels. However, there was no evidence that these patterns evolved as the NS participants' exposure to spoken English increased.

Experiment 2: The discrimination of English vowels

This experiment examined the NS participants' discrimination of English vowels. We used a categorial oddity discrimination test which required the 80 native Spanish and English participants to identify the serial position of an "odd item" out in change trials or else to press a button marked "no" for no-change trials presenting three physically different instances of a single vowel category (see, e.g., Guion et al., 2000; Lengeris & Hazan, 2007).

The contrasts between [i]-[ɪ], [ɑ]-[ʌ], [e^{ɪ}]-[ɛ], [æ]-[ɛ], [æ]-[ɑ], [ɪ]-[ɑ] and [i]-[u] stimuli were expected to yield scores ranging from near-chance to near-perfect. We expected high scores for [ɪ]-[ɑ] and [i]-[u] because these pairs of English vowels were seldom identified in Experiment 1 in terms of a single Spanish vowel (7% and 1% of the time, respectively). The amounts of "category overlap" observed for the other five contrasts were higher, ranging from 32% to 83%. It is not our intention, however, to predict discriminability from the cross-language mapping results because the aim of this experiment was to determine if the discrimination scores increased as a function of input.

Method

Stimuli

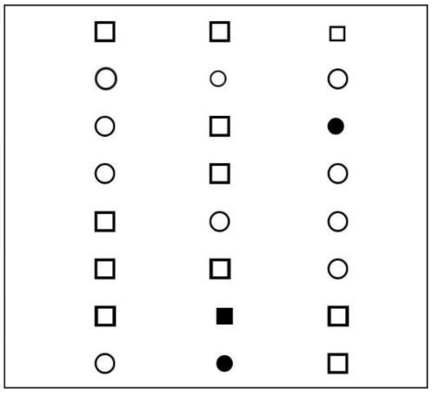
Five adult male native speakers of English from Birmingham, Alabama produced the English /i ɪ e ɛ æ ɑ ʌ u/ in a /bVbo/ context. The two syllables of these non-words were spoken with approximately equal stress.

The original tokens were edited to prevent variation in pre- and post-vocalic /b/ tokens from influencing perception of the target vowels. As expected (Dimitrieva et al., 2015, Figure 2), not all word-initial /b/ tokens were produced with pre-voicing. We therefore cross spliced the pre-voicing from one stimulus onto the remaining 39 stimuli. Next, the /o/ portions of all stimuli were removed at the point of complete constriction of the post-vocalic /b/. We then replaced the deleted post-vocalic portions of the stimuli with a single final [b] segment that was fully voiced and had an audible but low-amplitude release burst. The edited /bVb/ stimuli were then normalized for peak intensity.

Procedure

Each vowel contrast was tested by 10 change and 10 no-change trials. Each change trial contained two physically different realizations of one vowel category and a single realization of a different vowel category. No-change trials, on the other hand, consisted of three physically different realizations of a single vowel category. The three stimuli presented in each trial were separated by 1.0 sec. The participants' task was to identify the serial position ("1", "2" or "3") of the odd item out in change trials or else to push a fourth button, marked "no", in response to no-change trials.

Figure 5. Trials from the visual task used to familiarize participants with the categorial discrimination task used in Experiment 2. See text.



The discrimination task was introduced with a visual task that made use of geometric figures (circles and squares). As shown in Figure 5, some visual trials contained an odd item out whereas others did not. The participants were given no explanation regarding how to perform the visual task. They were simply asked to select one of four response alternatives ("1", "2", "3" or "no") from the answer sheet. Feedback (correct or incorrect) was provided after each response. The task familiarization phase continued until each participant gave eight consecutive correct responses. This was taken to mean that they had understood the nature of the oddity task, namely, that they were to identify the odd items out in change trials (which included, for example, one circle and two squares) but to respond "no" to no-change trials, ignoring variation in the size and/or color of the geometric figures.

Once criterion had been reached, the participants were told that the next task with vowels presented via headphones was "similar" to the task they had just successfully completed. They were given practice with feedback with some of the auditory stimuli described earlier. Each practice block consisted of five change and three no-change trials testing the contrast between English [e^{ɪ}] and [ɑ] stimuli. The practice block was repeated with different random orders until participants responded correctly to all eight trials in a practice set. All participants managed to do so within three blocks.

The discrimination test consisted of 70 change trials (7 vowel contrasts x 10) and 40 no-change trials (8 vowel categories x 5). The trials were randomly presented over headphones at a self-selected comfortable level, without feedback. The interval between the three stimuli in each trial was 1.0 sec. The inter-trial interval was also fixed at 1.0 sec. Trials could be replayed but responses could not be changed once given.

An A' score was calculated for each of the seven contrasts based on correct responses to change trials (counted as "hits") and incorrect selection of an odd item out in non-change trials ("false alarms"). The A' scores ranged from 0.5, indicating a lack of sensitivity, to 1.0, indicating perfect sensitivity (see Snodgrass et al. 1985).

Unlike the AXB discrimination test that is usually used to test the discrimination of foreign-language or L2 speech sounds, the categorial oddity test used here is subject to response bias (see Flege, 2003). Despite the familiarization procedures described earlier, six NS participants failed to obtain an A' score greater than 0.80 for [i]-[u]. This contrast served as a control inasmuch as the English [i] and [u] stimuli were consistently identified as instances of two distinct Spanish vowels, /i/

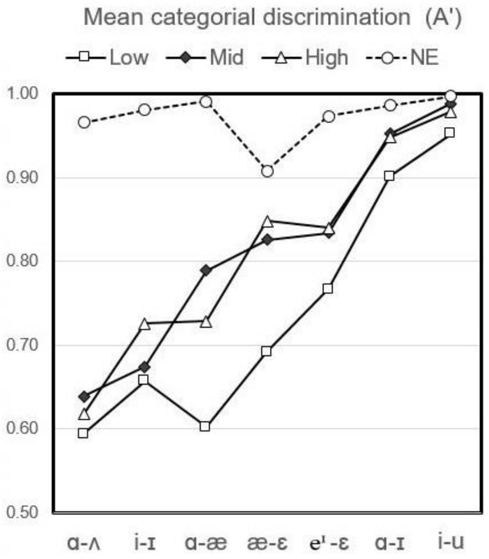
and /u/ (see Table 1). The low A' scores obtained for the excluded participants were due primarily to the incorrect selection of an odd item out in no-change trials (*M* = 72%). Despite the visual and auditory practice, these participants appear to have been biased to select on odd item out in response to *any* (even non-criterial) audible difference between the three physically different vowel stimuli in the no-change trials. The excluded NS participants were replaced by six drawn from the same population.

Results and Discussion

Figure 6 shows the mean A' scores obtained for the seven English vowel contrasts by the 20 NE participants and the 60 NS participants. Most NS participants obtained high scores for the control contrast, [i]-[u], and also for [ɪ]-[ɑ]. The high scores obtained for [i]-[u] and [ɪ]-[ɑ] might have been due, at least in part, to the development of new phonetic categories for the English vowels /u/, /ɪ/ and /ɑ/. We think it more likely, however, that the high discrimination scores obtained for the [i] and [u] stimuli arose from the generation of distinct phonetic codes associated with two different L1 categories, namely Spanish /i/ and /u/. Similarly, the high scores for the trials containing [ɪ] and [ɑ] stimuli may have been due to the generation of codes associated with the Spanish vowels /i/ or /e/ (for the English [ɪ] stimuli) and codes associated with a different Spanish vowel (/a/, /o/ or /u/) for the English [ɑ] stimuli.

We obtained lower scores for the remaining five contrasts. The primary reason for these lower scores, we think, is the lesser likelihood that the NS participants generated two distinct phonetic codes when processing stimuli that were, for the NE participants, phonetically distinct. However, our aim here is not to account for between contrast differences but to evaluate the role of input.

Figure 6. Mean sensitivity (A') to the contrast between seven pairs of English vowels for three groups of native Spanish adults (Low, Mid, High) and native English (NE) speakers. See text.



The A' scores were submitted to a (4) Group x (7) Vowel Contrast ANOVA with repeated measures on Vowel Contrast. It yielded significant main effects of Group, *F* (3, 76) = 33.0, *p* < .001, and Vowel Contrast, *F* (6, 456) = 43.8, *p* < .01, as well as a significant two-way interaction, *F* (18, 456) = 5.3, *p* < .001. The interaction was explored by tests of simple main effects followed by Tukey post-hoc tests with an alpha level of $p = .05$.

The simple effect of Group was significant for all seven vowel contrasts, F (3, 76) = 4.9-20.2, $p < .01$, whereas the simple effect of Vowel Contrast was significant for the three NS groups, F (6, 456) = 18.3-22.2, $p < .001$) but not for the NE group, F (6, 456) = 1.0, $p > .10$. The lack of a significant effect of Vowel Contrast for the NE speakers was expected given that they were likely to have generated two distinct phonetic codes when processing the stimuli used in testing all seven contrasts.

The post-hoc tests revealed that scores obtained from members of the Low input group were significantly lower for [i]-[u] and [ɑ]-[ɪ] than were the scores obtained for the Mid and High groups ($p < .05$). Perhaps, owing to their very limited exposure to spoken English, members of the Low group were less able than members of the Mid and High group to perceptually relate the English vowel stimuli to phonetically distinct Spanish vowels.

Lower scores were obtained for [e$^{\text{ɪ}}$]-[ɛ], [ɑ]-[æ] and [æ]-[ɛ] than for the two contrasts just mentioned. For [e$^{\text{ɪ}}$]-[ɛ], members of the Low, Mid and High groups all obtained significantly lower scores than NE speakers did ($p < .05$) but did not differ from one another. In Experiment 1, the members of all three NS groups tended to identify the [e$^{\text{ɪ}}$] and [ɛ] stimuli as instances of a single Spanish vowel, usually /e/. Moreover, Spanish /e/ may be realized as both an [e] and [ɛ]-quality vowels, as in the English words *ate* and *bet* (Dalbor, 1980, p. 152). The two-to-one cross-language mapping pattern, together with the allophony present in Spanish, seems to have prevented phonetic learning from occurring.

A two-to-one identification pattern also existed for the stimuli in [ɑ]-[æ] trials, which contrasted the low back English vowel /ɑ/ with the low front English vowel /æ/. Both the [ɑ] and [æ] stimuli were usually identified as instances of the low central Spanish vowel /a/ (see Table 1). Despite this, modest evidence of phonetic learning was obtained. The Low group obtained significantly lower scores for [ɑ]-[æ] than the Mid and High groups who, in turn, obtained lower scores than the NE group ($p < .05$).

The results for [ɑ]-[æ] recall the findings of a multi-dimensional scaling analysis by Fox et al. (1995). The aim of that study was to identify the perceptual features used by NS participants when perceiving English and Spanish vowels. The two NS groups tested by Fox et al. (1995) differed in years of L2 input (M = 0.9 vs 4.3 years, calculated here for the first time). The more experienced NS group was able to make greater use than the less experienced group of a perceptual dimension that distinguished central vowels, such as Spanish /a/, from front and back (i.e., non-central) vowels such as English /æ/ and /ɑ/. The authors interpreted this finding to mean that the NS participants' perceptual processing of vowels “changed gradually” as they gained experience with English (Fox et al., 1995, p. 2548).

For [æ]-[ɛ], members of the Low group obtained significantly lower scores than did members of the Mid, High and NE groups. Most importantly, neither the Mid nor High groups differed significantly from the NE group, which demonstrates phonetic learning.

The absence of a significant between-group differences for [æ]-[ɛ] between the NE group and NS participants having 1.2 and 3.0 years of input (Mid, High) was favored by the relatively low scores obtained for [æ]-[ɛ] by the NE speakers. The less-than-perfect scores obtained by the NE-speaking listeners was likely due to an ongoing change in how front vowels are produced in Southern cities such as Birmingham, Alabama. In this dialect of American English, the vowel /æ/ is being produced with higher first formant (F1) frequency values than before. In effect, realizations of /æ/ are beginning to invade the portion of vowel space occupied by /ɛ/ (Labov et al., 2006).

The phonetic learning seen here for [æ]-[ɛ] is unsurprising. In Experiment 1 we saw that the members of all three NS groups tended to label the [æ] stimuli as Spanish /a/ and the English [ɛ] stimuli as Spanish /e/. And, as reviewed by Flege (1991b, p. 703), NS learners of English tend to realize the English vowels /æ/ and /ɛ/ as [a]-quality and [e]-quality vowels, respectively.

One possible explanation for the higher scores obtained for [æ]-[ɛ] by the Mid and High groups than by the Low input group is that the NS participants having more than about one year of spoken English input had begun establishing a new phonetic category for English /æ/. Such a development is envisaged by the Speech Learning Model (Flege, 1995). However, if this were true, then members of Mid and High should have given more "not Spanish" responses to the English [æ] stimuli than did members of Low input group. This finding was not obtained in Experiment 1, however, and so we think that a more likely explanation for the between-group difference for [æ]-[ɛ] was that members of the Mid and High groups were simply more effective in using Spanish vowels to identify the English [æ] and [ɛ] stimuli than were the NS participants who had just 0.2 years of English input (i.e., members of the Low group).

Finally, the lowest scores obtained in the present experiment were those for trials testing the [i]-[ɪ] and [ɑ]-[ʌ] stimuli. All three NS groups obtained significantly lower scores for these two contrasts than the NE group did ($p < .05$) and did not differ significantly from one another.

In summary, this discrimination experiment provided limited evidence of phonetic learning for English vowels. Members of the Low input group obtained significantly lower scores for [ɑ]-[æ] than did members of the Mid and High groups who, in turn, obtained lower scores than the NE group. For [æ]-[ɛ], the Low input group obtained significantly lower scores than the Mid, High and NE groups and, most importantly, neither the Mid nor the High input groups differed significantly from the NE group.

Experiment 3: Production of English vowels

This experiment examined the effect of L2 input on the production of eight English vowels. Production accuracy was evaluated via listener judgments (see below) rather than acoustically. The procedure we used to elicit English vowel production by the NS participants was explicitly designed to avoid unwanted effects of orthography.

Orthography is known to affect the production of English phonetic segments by native speakers of languages having a transparent relation between letters and pronunciation. Two such languages are the Romance languages Italian and Spanish. Bassetti (2017) found that young Italians who were studying English at a high school in Rome, Italy produced English consonants differently based on spelling. For example, the students produced the medial /p/ in words like *copy* with shorter stop closure durations than NE speakers did whereas they produced longer closure durations then NE speakers in words like *floppy* where the medial /p/ is spelled with two letters.

Flege, Bohn and Jang (1997) examined English vowel production by 20 NS adults living in Birmingham, Alabama. The vowels of interest were those found in the words *beat* ([i]), *bit* ([ɪ]), *bet* ([ɛ]) and *bat* ([æ]). The NS speakers inserted these and other test words they had read from a written list into an invariant carrier phrase (*Now I say_*). Two NS participants produced the /i/ in *beat* with an [e]- or an [ɛ]-quality vowel, suggesting that they produced the target English vowel /i/ not as it sounds but as the letter <e> is normally pronounced in Spanish. Even more NS participants produced the vowel in *bit* ([ɪ]) as an [i]-quality vowel. This may of course have been due to an inability to produce English /ɪ/ correctly, but at least some production errors might have been "spelling" pronunciations

(Bassetti, 2008).

The “spelling pronunciation” interpretation of the vowel production results obtained by Flege et al. (1997) was reinforced by the results obtained for the same NS participants in a vowel identification experiment. The NS participants and NE controls identified members of a synthetic continuum ranging from *beat* [bit] to *bit* [bɪt] by pressing buttons orthographically labelled <beat> or <bit>. Several NS participants reversed the expected labelling pattern, pressing the button marked <bit> (phonetically [ɪ]) to identify [i]-quality vowels and the button marked <beat> (phonetically [i]) to label [ɪ]-quality vowels.

Method

The NE participants produced 58 picturable English words. Pictures representing the 58 target words were presented in a random order along with a written version of the Spanish translations of the word or action represented in each picture. The NS participants were usually able to identify the intended English target words but, if they did not, the experimenter (R.W.) gave them additional cues. For example, if a participant said “kick the ball” instead of “hit the ball” the experimenter said “No, not kick, another word”. All items were produced twice to ensure fluency. The choice of words selected for analysis here was based on the number of fluently produced tokens that we managed to elicit plus a consideration of the ease or difficulty of editing them (see below).

The 15 English words eventually selected for analysis were digitized at 22.05 kHz. Whenever possible, the second production was used. Productions of the following vowels were examined: /æ/ in *bag* and *apple*, /e^{ɪ}/ in *baby*, *pay*, /u/ in *two*, *shoe*, /ʌ/ in *bus*, *but*, /i/ in *eat*, *key*, /ɪ/ in *big*, *sick*, /ɑ/ in *dog*, *sock;* and /ɛ/ in *bed*. Although we elicited four words with /ɛ / (*bed*, *red*, *ten*, *very*) mispronunciation of the /ɹ/ in *red* and *very* and variation in amount of nasalization in *ten* ruled out the selection of these words for analysis.

The selected test words were edited to reduce the possibility that consonant production errors might influence listeners’ judgments of the target vowels. Pre-vocalic tokens of /b/ and /d/ were produced with varying amounts of pre-voicing and, occasionally, with short-lag VOT. Accordingly, any pre-voicing present in these pre-vocalic consonants was digitally removed. To obviate an effect of VOT differences in *pay*, the burst and aspiration of each participant’s production of *pay* was replaced by the burst and short-lag VOT portions of the same participant’s production of *bay*. For the pre-vocalic consonants in *shoe*, *sick*, *sock*, *key* and *two*, the original consonant produced by each participant was replaced by a single token of the same consonant as produced by a NE speaker. Finally, all portions of the test words following the constriction of the post-vocalic consonants were removed.

The edited tokens were checked for quality by the two authors. The few stimuli they judged to not be natural-sounding were declared missing. The remaining tokens were then normalized for peak intensity. Of the 900 target words produced by the NS participants (15 words x 60 participants), 33 were declared missing because a NS participant was unable to determine the target word we wanted to elicit. A total of 102 other words were declared missing because the mispronunciation of a flanking consonant could not be corrected by editing (e.g., a spirantized /d/ or /g/ in *dog*) or because an edited token was judged to not be natural-sounding.

Auditory evaluation

Eight adult monolingual native speakers of English, mainly from the Southeastern portion of the United States, evaluated the 765 vowel tokens. All passed a pure tone hearing screen at octave frequency between 250-4,000 Hz.

The 15 target words were presented in separate counter-balanced blocks. The NE listeners were asked to judge the accuracy of the one target vowel presented in each block. Prior to beginning a block, the listeners were shown three written key words illustrating the target vowel under examination (e.g., *cat*, *bad* and *sad* for blocks with /æ/). The listeners were asked to say the key words aloud and to use their own pronunciation as a point of reference when judging the target vowels they would hear.

The listeners were told to judge each vowel using one of the following four response alternatives: “good”, “acceptable”, “distorted” or “wrong vowel”. They were told to press the “good” button (response alternative #1) if they heard a vowel corresponding closely to their own pronunciation; to use the “acceptable” label (#2) for stimuli that were readily identifiable as an instance of the target vowel but which differed slightly from their own pronunciation of it; and to use the “distorted” button (#3) for recognizable vowels differing substantially from their own pronunciation of the target vowel. Use of the “wrong vowel” response alternative (button #4) was reserved for vowel stimuli that listeners judged to *not* be an instance of the target vowel.

Listeners could replay stimuli if they wished but could not change a response once given. The next trial was presented 1.0 sec later. The stimuli in each block were randomized differently for each listener. Ten items from the end of the randomized list were duplicated at the beginning of the block. Responses to these trials were treated as practice and so not analyzed.

The four response alternatives were not treated as a scale. Instead, we derived two variables for productions of the eight English target vowels. The “Percent Correct” scores indicated the percentage of times that the NE-speaking listeners judged each participant’s production of a target vowel to be an instance of the intended category (i.e., when response alternatives #1-3 were selected). The “Percent Good” scores, on the other hand, indicated *non-distorted* productions of the target vowels, that is, the percentage of times that response alternative #1 or #2 was selected.

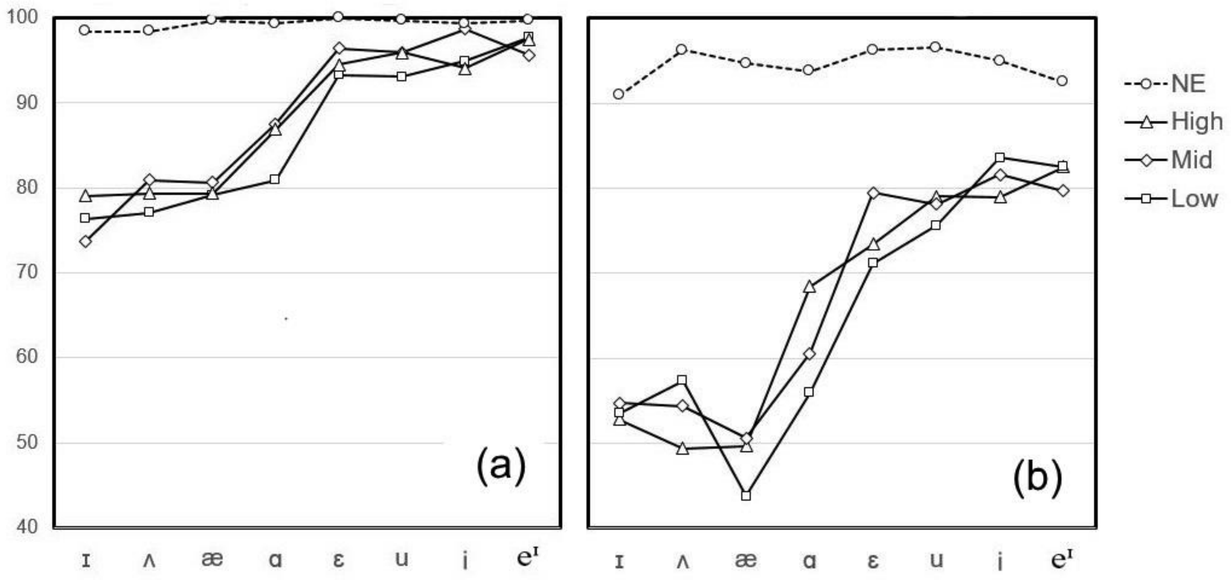
Result and discussion

The mean Percent Correct scores are shown in Figure 7(a). The English vowels were judged to have been produced in a recognizable fashion 99.3% of the time, on average, by the NE-speaking listeners. The Percent Correct scores obtained for the 60 NS participants averaged 87.9%, ranging from 76% for the target vowel /ɪ/ (in the words *big* and *sick*) to 96.9% for /eɪ/ (in *baby*, *pay*).

The missing data mentioned earlier resulted in unequal number of Percent Correct scores for the eight target vowels. We therefore tested the effect of Group (4 levels) separately for each target vowel. When the effect of Group reached significance we tested for pairwise between-group differences using Tukey post-hoc tests (alpha level .05).

The effect of Group did not reach significance for the target vowels /ɛ/, /u/, /i/ and /eɪ/. However, all three groups of NS speakers received significantly lower Percent Correct scores than the NE group did for /ɪ/, /ʌ/ and /æ/ (p < .05). For these English vowels there were no significant differences between the three NS groups. For /ɑ/, only Low differed significantly from the NE group. Although this indicates phonetic learning by members of the Mid and High groups, it is worth noting that the differences in Percent Correct scores obtained for the Early (M = 80.9%) and the Mid and High groups (M = 87.5% and 86.9%, respectively) were small.

Figure 7. The accuracy with which eight English vowels were produced by native English (NE) speakers and native speakers of Spanish differing in L2 input: High, Mid and Low. (a) The percentage of vowels identified as an instance of the intended target vowel category; (b) The percentage of non-distorted realizations of the intended target vowel.



The mean Percent Good scores are shown in Figure 7(b). Scores obtained for the NE group averaged 94.5%, which is 4.8% lower than the Percent Correct scores obtained for the NE speakers. This means that 4.8% of the vowels produced by the NE speakers were judged by NE-speaking listeners to have been “distorted” instances of their intended categories. This was likely to have been due to dialectal differences between the individual listeners who judged the vowels and the vowels spoken by some of the 20 NE speakers.

The Percent Good scores obtained for the 60 NS speakers averaged 66.5%, substantially lower than their Percent Correct scores. The difference between the Percent Correct and Percent Good scores indicates that 21.4% of the NS participants’ production of the English vowels were judged to be recognizable but distorted instances of the intended English vowel categories.

Significantly lower Percent Good scores were obtained for /ɪ/, /ʌ/, /æ/, /ɑ/ and /u/ by all three NS groups than productions of these vowels by the NE speakers (*p* < .05). For these vowels there were no significant differences between the High, Mid and Low input groups. For /ɛ/, on the other hand, only the Low and High groups differed significantly from the NE group (*p* < .05).

No significant differences between the three NS groups were noted for /i/ and /eᶦ/, nor between any NS group and the NE speakers. The results for /i/ probably do not demonstrate phonetic learning. This is because the difference between the /i/ vowels of English and Spanish are very small both in articulatory and acoustic terms (Flege, 1989). Moreover, these small cross-language differences for /i/, even if they are auditorily detectable, are probably obscured by differences in how NE adults living in Birmingham perceive English /i/. Intra-listener differences among native speakers of English in Birmingham are nearly as great as differences in how /i/ is produced in English and Spanish (Frieda et al. 1999).

The lack of significant native vs non-native differences in the production of the target vowel /eᶦ/, on the other hand, may potentially indicate phonetic learning. Like Spanish /e/ (Flege, 1989), the /e/ of Italian is produced with very little formant movement. Flege et al. (2003) evaluated L2 vowel production accuracy using the same procedures used in the present experiment. A substantial amount of variance in the accuracy with which native Italian speakers were judged to have produced English /eɪ/ was accounted for by amount of acoustically measured formant change. Given that the NE-speaking listeners in the Flege et al. (2003) study were also from Birmingham, Alabama, we think that the NE-speaking listeners in the present study would likely not have judged the NS talkers' productions of the target vowel /eɪ/ as "good" unless these productions had been produced with a sufficient amount of formant movement.

However, given that even productions of English /eɪ/ by member of the Low group were accepted by the NE-speaking listeners, an alternative explanation for the seemingly accurate production of /eɪ/ is likely. Having detected differences between English /eɪ/ and Spanish /e/ in terms of amount of tongue movement (Flege, 1989), the NS participants in the present study may simply have re-interpreted the English vowel /eɪ/ as a "fused" sequence of the Spanish vowels /e/ and /i/ (Dalbor 1980, p. 182), as in the Spanish word *seis* ("six").

In summary, all three NS groups produced /ɪ ʌ æ ɑ u/ less accurately than the NE speakers did. For a sixth vowel examined, /ɛ/, this held true for two of the three NS groups. For these six vowels there were no significant differences between the three NS groups. For /i/ and /eɪ/, on the other hand, none of the NS groups differed significantly from the NE group. The results for /i/ probably do not indicate that phonetic learning had occurred given that the unmodified use of Spanish /i/ in English words is likely to be readily accepted by NE-speaking listeners. Nor can we be sure that an improved production of /eɪ/ demonstrated phonetic learning. It might have been due to the use of a different Spanish vowel as a substitute for the target English vowel.

Experiment 4: Consonant discrimination

Our final experiment evaluated the effect of L2 input on the discrimination of English consonants. We used the categorial oddity test from Experiment 2 to examine the contrasts between [d̥]-[ð] (as in *dough* vs *though*), [b̥]-[v] (as in *ban* vs *van*), [s]-[z] (*sap*, *zap*), [θ]-[ð] (*thatch*, *that*), [s]-[θ] (*sick*, *thick*), [d̥]-[tʰ] (*dot*, *taught*) and [b̥]-[s] (*boss*, *sauce*).

The [b̥]-[s] contrast served as a control. The English [b̥] stimuli were produced without pre-voicing or aspiration and so closely resemble short-lag realizations of Spanish /p/ in words like *peso*. The English alveolar [s] stimuli resembled the initial dental fricative in Spanish words like *sol* ("sun"). We reasoned that the NS participants would be able to discriminate the [b̥] and [s] stimuli at high rates, even in the absence of phonetic learning, simply by generating the phonetic codes associated with Spanish /p/ and /s/ when processing the English [b̥] and [s] stimuli.

In the absence of cross-language perceptual mapping data like that provided earlier for vowels (Experiment 1) we are unable to offer theory-based predictions regarding how easy or difficult the remaining six consonant contrasts would be for our NS participants. However, as for the vowel discrimination experiment reported earlier we expected to observe a wide variation in perceptual sensitivity to the consonant contrasts.

Method

Five adult male NE speakers produced non-word disyllables by inserting /b̥ d̥ t s v θ ð/ into a /'Cɑdo/ frame. The /b̥/ and /d̥/ tokens were produced without pre-voicing and so had the voice onset time values typical for Spanish /p/and /t/, respectively. After all portions of the signal following complete constriction of the intervocalic /d/ had been removed, the edited CVC stimuli were normalized for peak intensity.

The [d̥]-[ð], [b̥]-[v], [s]-[z], [θ]-[ð], [s]-[θ], [d̥]-[t^h] and [b̥]-[s] contrasts were each tested by 10 change and 10 no-change trials. The interval between the three stimuli in each trial was 0.5 sec; the interval between each response (a button marked “1”. “2”. “3” or “no”) and the following trial was 1.0 sec. The listeners’ sensitivities to the seven phonetic contrasts were indexed by A’ scores based on the proportion of hits (correct identifications of the odd item out in change trials) and false alarms (incorrect selections of an odd item out in no-change trials).

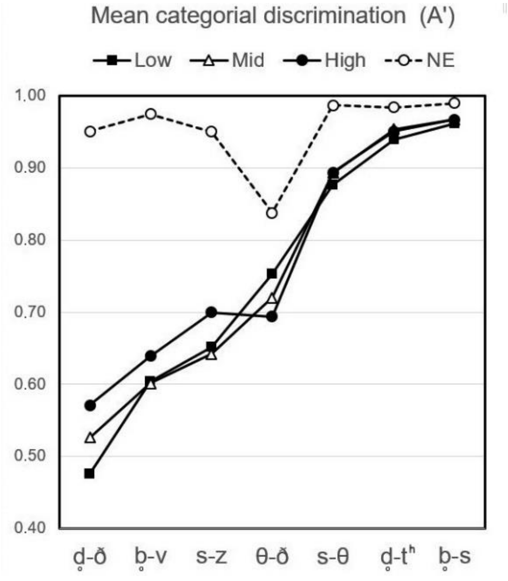
The visual practice used for vowel discrimination in Experiment 2 was not repeated here. However, all participants were required to respond correctly to all eight items in a practice block before beginning the consonant discrimination test. The practice blocks consisted of five change and three no-change trials testing the contrast between [d̥]-[s] using stimuli described earlier. Feedback was provided during the practice session, but not during the test. The stimuli in a trial could be replayed but responses could not be changed once given.

Results and discussion

The mean A’ scores obtained for the four groups are shown in Figure 8. The NE group obtained very high scores for all contrasts except the one between the voiced and voiceless inter-dental fricatives [ð] and [θ]. The NE group’s less than perfect scores for [θ]-[ð] came as no surprise given that NE-speaking listeners often confuse [θ] and [ð] in an identification task when these fricatives occur in CV syllables embedded in noise (e.g., Cutler et al., 2004). NE speakers’ perceptual difficult with the [θ]-[ð] contrast exists because the two English word-initial fricatives are brief in duration and low in amplitude and, moreover, carry low functional load.

As expected, the members of all three NS groups obtained high scores for the control contrast [b̥]-[s]. Also as expected, their scores for the remaining six contrasts varied considerably, ranging from near-chance to near-perfect.

Figure 8. Mean sensitivity (A’) to the contrast between seven pairs of English consonants by members of three native Spanish groups (Low, Mid, High) and a group of native English (NE) speakers.



The A' scores were examined in a (4) Group x (7) Consonant Contrast ANOVA with repeated measures on the Contrast factor. The ANOVA yielded significant main effects of Group and Contrast as well as a significant two-way interaction, F (18, 456) = 10.1, p < .0001. The interaction was explored by simple effects tests followed by Tukey post-hoc tests with an alpha level of .05.

The most important finding was the lack of significant differences between the NS groups differing in English language input. We obtained significantly lower scores for the Low, Mid and High input groups than for the NE group for [d̥]-[ð], [b̥]-[v], [s]-[z] and [s]-[θ]. For a fifth contrast, [θ]-[ð], only the High group differed significantly from the NE group. For the remaining two contrasts, [d̥]-[t^{h}] and [b̥]-[s], no NS group differed from the NE group.

The effect of Contrast was highly significant for all four groups (F-values ranging from 4.1-48.4, p < .001). The NE participants' scores were significantly lower for [θ]-[ð] than for the other six contrasts. For the Low and Mid groups, scores for [d̥]-[ð] and [b̥]-[v] were significantly lower than those obtained for [s]-[z] and [θ]-[ð], and the scores for these contrasts were significantly lower than those for [s]-[θ], [d̥]-[t^{h}] and [b̥]-[s] (p < .05). For the High group, significantly lower scores were obtained for [d̥]-[ð], [b̥]-[v], [s]-[z] and [θ]-[ð] than for [s]-[θ], [d̥]-[t^{h}] and [b̥]-[s].

The significant between-contrast differences obtained for the NS participants might be explained, at least in part, by considering the frequency with which the various English consonant stimuli used here can be heard on the phonetic surface of Spanish. Purely acoustic factors are likely to play a role as well (Cutler et al., 2004). However, given that the matter is complex, and given that attempted explanations that might be offered here would be speculative at best, we will not enter into a discussion of the between-contrast differences but will instead focus on the unexpectedly high scores obtained by the NS participants for the [d̥]-[t^{h}] contrast.

The high [d̥]-[t^{h}] scores obtained for all three NS groups might be interpreted to mean that they had established new phonetic categories for English /t/, which is realized with long-lag VOT ([t^{h}]) rather than with the short-lag VOT used for Spanish /t/. We think it more likely, however, that the high scores resulted from the application of a response strategy rather than the development of a new phonetic category for English.

Of the 60 NS participants tested, 28 obtained perfect A’ scores (1.0) for [d̥]-[t^{h}] and 27 others made two or more errors, obtaining A’ scores ranging from 0.67 to 0.95 (M = 0.89). We carried out ANOVAs to identify differences between the subgroups of NS participants who obtained perfect scores (“P”) and non-perfect scores (“NP”). The P and NP subgroups did not differ significantly from one another in age of arrival in the United States (M = 28.9 vs. 26.6 years), length of residence in the U.S. (M = 2.4 vs. 2.5 years), self-reported percentage use of English (M = 58.1% vs 51.2%) or years of English input (M = 1.5 vs 1.3 years). However, members of the P subgroup had studied English significantly longer in school than members of the NP subgroup did (M = 8.0 vs 3.6 years, F (1, 53) = 24.1, p = .0001).

Formal classroom instruction rarely if ever impacts the phonetic details of L2 segmental production and perception. The present finding appears to be an exception. We think that the NS participants who obtained perfect scores for [d̥]-[t^{h}] were more likely than those who obtained somewhat lower scores to have learned in school that English /t/ differs from Spanish /t/ in the duration of post-release aspiration (i.e., VOT). The formal instruction in English that members of the P subgroup had obtained prior to immigrating to the United States may have enabled them to obtain perfect scores for [d̥]-[t^{h}] by generating the phonetic code associated with Spanish /t/ when processing the [d̥] stimuli and by coding the [t^{h}] stimuli as “NOT Spanish”. Use of an X-not X strategy, if that is what enabled some NS participants to obtain perfect [d̥]-[t^{h}] scores, is quite different from generating two distinct *phonetic* codes, one associated with Spanish /t/ and one associated with a new phonetic category English /t/, when processing the [d̥] and [t^h] stimuli. We will return to this point in the next section.

In summary, the NS participants readily discriminated the control contrast [b̥]-[s] but showed a wide range of scores for the other six English consonant contrasts examined here. The lack of a significant difference between the three NS groups suggests that an increase in spoken L2 input from 0.2 to 3.0 years did not result in a better discrimination of English consonants.

General Discussion

We evaluated the effect of quantity of L2 input by comparing three groups of native Spanish (NS) adults differing in years of full-time English input (Low 0.2, Mid 1.2 and High 3.0 years). All were Late learners who spoke English with strong foreign accents. There was little evidence that an increase in input from 0.2 to 3.0 years was sufficient to change the perceived relation between Spanish and English vowels (Experiment 1), the discrimination of English vowels (Experiment 2), accuracy in producing English vowels (Experiment 3), or the discrimination of English consonants (Experiment 4). Apart from a modest improvement in vowel discrimination, our examination of input differences in Late learners who were, overall, relatively inexperienced in English yielded what is essentially a null finding.

The present study, like many other studies of naturalistic L2 learning, yielded fairly clear results that cannot be adequately explained. Two very different explanations might be offered for the limited effect of input variations observed here. One is that *no amount* of L2 speech input that might be received by "post-Critical Period" learners will enable them to perceive, and eventually produce the phonetic segments of an L2 in a fully native-like fashion (see, e.g., Long, 1990; Granena & Long, 2012; Hyltenstam & Abrahamsson, 2003; DeKeyser, 2000).

This explanation is contradicted by results for Late learners that are evident in Figure 1, but for some the "maturational" account just offered might seem to be an unrealistic portrayal of the critical period hypothesis derived from Lenneberg's (1967) original formulation. However, it is not unrealistic. Lenneberg suggested that post-Critical Period learners may no longer be able to make "*automatic*" use of input "*from mere exposure*" and that the success post-Critical Period learners might enjoy in an L2 comes at the cost of "*conscious and labored effort*" not needed for successful L1 speech acquisition (Lenneberg, 1967, p. 176). All of the Late learners in Figure 1 might have suffered the ill effects of having passed a critical period but only those who possessed special aptitude for L2 learning (e.g., unusually great phonological short-term memory, see MacKay et al., 2001, and references therein) or had worked especially hard to learn English pronunciation (perhaps due to unusually strong motivation) managed to produce English /t/ in a native-like fashion.

An implicit assumption underlying a maturational account explanation of L2 speech learning is that some basic capacity used in establishing the L1 phonetic system and that remains available until puberty is lost or attenuated following the closure of a critical period at puberty. The loss or attenuation of capacity for speech learning might potentially be restricted to one or more of the following more basic capacities: the ability to detect cross-language phonetic difference; the ability to store and structure detected cross-language differences in long-term sensory memory; the ability to transform sensory based perceptual representations into motoric codes needed for speech production (see Flege, 1988, for a comprehensive early review of these issues).

A second possible explanation of the null finding obtained in the present study begins with a rejection of the critical period construct as well as the associated assumption of a reduced or attenuated capacity for speech learning (see, e.g., Flege, 1987, 1995). On this view, non-nativeness in L2 segmental production and perception might be attributed in part to the continued influence of L1 phonetic elements, but primarily to the inadequate input received by many learners of an L2 (Flege, 1987, 1995, 2008, 2018). On this view, the seeming lack of phonetic learning in the present study was not due to a diminished capacity for learning speech after puberty but to *inadequate input*.

The competing explanations just offered, of course, reflect a far more general debate between empiricists and nativists who view the role of input quite differently (e.g., Bates & Elman, 1996; Cowie, 1999; Sampson, 2005; Tomasello, 2009). We think it fair to say, however, that most SLA researchers are "empiricists" with respect to the nature of native-language speech learning and that most (possibly all) would agree that the slow process of phonetic learning in the L1, which eventually enables monolingual children to "sound" like those around them, is the result of a successful "attunement" to speech input (Aslin & Pisoni, 1980). A crucial question for L2 speech research, then, is whether *all* learners of an L2 are equally capable of successful attunement to the L2 phonetic system, or whether such a capacity is limited to "pre-Critical Period" learners.

A clear response to this question is needed. Otherwise it is difficult to know when observed instances of non-nativeness in L2 segmental production and perception is the result of *permanent limits* on the capacity to learn speech and when observed instances of non-nativeness is simply a manifestation of "learning in progress". To help better frame this question for non-specialists, we will now briefly review some data pertaining to learning of the VOT dimension in an L2. We then conclude the article by sketching a new experimental technique that might be used to choose between the competing explanations just outlined.

Learning English /p t k/

Native Spanish (NS) learners of English must learn to produce word-initial tokens of English /p t k/ with long-lag VOT values rather than with the short-lag VOT values used for Spanish /p t k/. To do so reliably in conversational speech, they must establish new phonetic categories for English /p t k/.

If the process of learning L1 and L2 speech is the same, as maintained by Flege (1995), NS learners of English will need *as much* native-speaker input to establish new phonetic categories for English /p t k/ as monolingual children do. How much input is that? Developmental research indicates that monolingual children exposed only to native-produced English need *at least* 10 years of full-time input to produce and perceive English /p t k/ like monolingual English adults (e.g., Eguchi & Hirsch, 1969; Elliott, 1979; Elliott et al., 1986; Flege & Eefting, 1986; Hazan & Barrett, 1999; Lee et al., 1999; Johnson, 2000; Koenig, 2001, Whiteside & Dobbin, 2003). Less research is available for L1 Spanish development. However, Flege and Eefting (1986) found that 8-9 year-old monolingual Spanish children differed significantly from monolingual Spanish adults in producing and perceiving the VOT dimension in word-initial stops.

The production and perception of English /p t k/ by NS learners of English has been studied extensively and so only a few studies can be cited here. The first study to directly compare Early to Late learners was that of Flege (1991a). This study tested NS adults in Austin, Texas who began learning English before or after puberty. The mean VOT values produced by Early learners were virtually identical to those of English monolinguals whereas the mean VOT values obtained for a group of Late learners was roughly equidistant from the mean values obtained for Spanish and English monolinguals. Flege (1991a) suggested that the inaccurate production of English stops by the Late learners was due to the fact that, having begun to learn English after puberty, they were unable to establish new phonetic categories for English /p t k/.

In retrospect, we see another possible explanation for non-nativeness in the Late learners' production of English stops. Unlike the Early learners, the Late learners in Texas had not yet received 10 years of native-speaker input. In fact, inspection of the mean values obtained by Flege (1991a) for individual Late learners reveals that some seemed to have learned nothing at all, producing English /t/ as if it were Spanish /t/, whereas other produced English /t/ with VOT values like those of native English speakers. In fact, the range of VOT values observed for the Late learners in Texas resembled those obtained for Late learners in Birmingham, Alabama by Flege et al. (1998) and shown here in Figure 1.

In Experiment 4 of the present study we found that all three NS groups obtained high discrimination scores for the contrast between short-lag and long-lag English stops ([d̥] vs [tʰ]). This finding might be interpreted to mean that the Late learners, even those with very little exposure to spoken English, managed to establish new phonetic categories for the long-lag [tʰ] of English. This is unlikely, however, given the evidence cited earlier regarding how much input is needed by monolingual English children to do so. And, as we noted when presenting the Experiment 4 results, years of formal instruction in English exerted a strong effect on the Late learners' discrimination of English [d̥] vs [tʰ] stimuli. This led us to conclude that the Late learners obtained high discrimination scores by applying an "X vs Not-X" decision strategy, not by generating two distinct phonetic codes for the English [d̥] and [tʰ] stimuli.

Why does the formation of phonetic categories take so long in L1 acquisition (and, by extension, in L2 learning)? Phonetic categories are *perceptual* representations that speaker-hearers establish over time in long-term memory based on the phonetic input they receive. A phonetic category is defined by *all* of the tokens encountered on the phonetic surface of meaningful speech that have been identified as instances of the category. The details specified by language specific phonetic categories are important for word recognition (van Alphen & Smits, 2004; van Alphen & McQueen, 2006), and also because they guide the development of the language specific realization rules used in speech production.

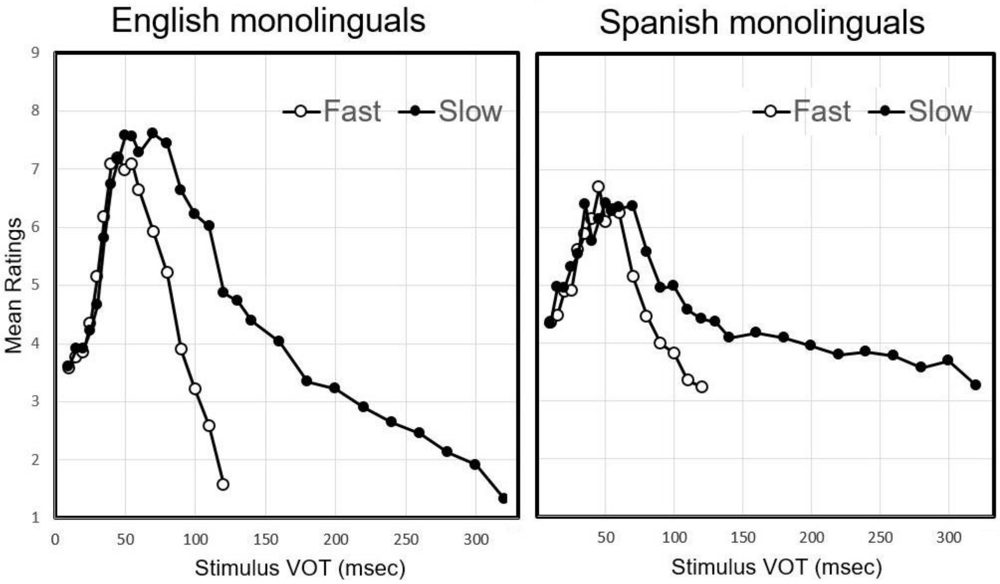
According to Flege and Schmidt (1995, pp. 92-93) phonetic categories are inherently multidimensional. Part of the process of establishing phonetic categories involves specifying how multiple acoustic dimensions (or "cues") are integrated and weighted. For English /p t k/ in word-initial position, these dimensions include not only VOT but also F0 frequency, F0 changes, F1 onset frequency and subsequent changes, aspiration intensity, and burst intensity and frequency. Flege and Schmidt (1995) also noted that the absolute normative value for the various dimensions, as well as their relative importance, may vary as a function of "phonetic context … degree of stress or emphasis, and speaking rate" (pp. 92-93).

The development of language-specific phonetic categories permit speaker-hearers to know how realizations of /p t k/ "ought" to sound in a particular speech variety. Given that phonetic categories guide the development of phonetic realization rules, mature speaker-hearers of a specific language variety will eventually re-produce consonants such as /p t k/ with the acoustic phonetic properties present in the input distributions to which they have been exposed. Individual differences in the L1 phonetic categories of monolinguals may, of course, arise from exposure to somewhat different input distributions (see, e.g., Frieda et al., 1999) as might be seen for the VOT dimension in studies examining monolinguals differing in dialect (e.g., Docherty et al., 2011). By extension, the same will also hold true in the L2 according to the Speech Learning Model proposed by Flege (1995)

Flege, Schmidt and Wharton (1996) examined the perception of English stops by English monolinguals, Spanish monolinguals, and NS Early and Late learners of English. The results of this study can be used to illustrate some of the just mentioned properties of phonetic categories. The stimuli used in two experiments by Flege et al. (1996) were synthetic /bi/-/pi/continua in which stimulus VOT varied from short-lag to very long-lag (i.e., VOT values longer than is usual for conversational speech spoken at a particular rate). One continuum consisted of CVs having relatively long vowels, simulating a relatively slow speaking rate; the other consisted of CVs having shorter vowels, simulating a faster speaking rate. Speaking rate was manipulated because VOT varies systematically as a function of speaking rate in the production of English but not Spanish.

In an identification experiment, English monolinguals used buttons marked "b", "p" and "exaggerated p" to label members of the two VOT continua (presented in counterbalanced order). English monolinguals identified a well-defined range of stimuli as "English p". They showed a clear cross-over from "b" responses (for stimuli with short-lag VOT) to “p” responses to "exaggerated p" responses (for stimuli having VOT values that were too long for normal-rate English stops). The Spanish monolinguals used buttons marked "Spanish p", “English p" and "exaggerated p" to identify the same stimuli. They were, of course, unable to identify a well-defined range of stimuli as “English p” because they did not possess phonetic categories for long-lag English stops.

Figure 9. The mean goodness ratings obtained by Flege et. al. (1996) from English and Spanish monolinguals for two VOT continua simulating relatively fast and slow speaking rates. See text.



Participants in the Flege et al. (1996) study then rated members of the two VOT continua for “goodness” as instances of the English /p/ category using a 9-point scale. As shown in Figure 9, English monolinguals’ ratings of normal-rate stimuli first increased as the VOT values began to resemble those typical for normal rate speech, then systematically decreased as VOT values increased beyond typical values. An increase-decrease pattern was also observed for the slow rate VOT continuum, but with the entire distribution shifted to the right, towards longer VOT values. The rate-induced perceptual shift was in keeping with the fact that in English (but not Spanish) VOT increases in word initial stops when speaking rate slows. Also as seen in Figure 9, Spanish monolinguals used a restricted range of ratings because they were unable to define appropriate VOT parameters for English

/p/. As expected, they also showed a significantly smaller shift due to the speaking rate manipulation than the NE monolinguals did.

The goodness ratings obtained for all ten Early learners closely resembled the pattern seen in Figure 9 for English monolinguals. This was interpreted to mean that all of the Early learners had established phonetic categories for English /p/ (see also Flege & Eefting, 1988). Two different patterns were observed for the Late learners, however. The four (of ten) Late learners who showed the pattern seen in Figure 9 for Spanish monolinguals produced English stops with Spanish-like short-lag VOT values. The four late learners who showed the pattern seen in Figure 9 for English monolinguals, on the other hand, produced English /p/ with English-like long-lag VOT values, presumably because they had established new phonetic categories for English /p/.

Why the difference between the two groups of four NS Late learners each? We think that this difference was unrelated to maturational state but arose, instead, from input differences: adequate for the "successful" but not for the "unsuccessful" Late learners.

Evaluating the role of input in L2 speech learning

From a methodological standpoint, L2 speech research is currently at a standstill. Piske and Young-Scholten (2008) observed that "we do not know how much input second language learners actually get [nor] how much exposure a learner requires" (2008, p. 13). The same authors noted that the input received by some adult immigrants may be "severely limited" (see also Moyer, 2008) and that a typical adult immigrant is unlikely to receive "anything comparable to the "9000 hours of input a child has received by the age of five" (2008, p. 13).

We think that new measures of the quantity and quality of L2 input will be needed to evaluate the extent to which speech learning in an L2, like that which takes place during L1 acquisition by monolingual children, is input driven. As mentioned in the Introduction, the current practice in L2 speech research is to use written language background questionnaires (LBQs) to estimate percentage L2 use. Participants are asked, for example, to self-estimate their overall use of the L2 in various time (e.g., currently, for the past 5 years). This approach is inadequate because, as far as we know, the percentage use estimates obtained from LBQs have never been externally validated. Just as importantly, LBQs provide little or no insight into how much L2 input is *foreign accented*.

Moyer (2008) suggested that a better understanding of L2 input might be drawn from more detailed "introspective reports and interviews" (2008, p. 173). While this would surely be useful, we think that a fundamentally new approach is needed. The approach we have in mind was first identified for L2 speech learning by Flege (2008, p. 189). It involves using the Experience Sampling Method (ESM) developed by Csikszentmihalyi and Larson (1987) to quantify both the quality and quality of L2 input.

The ESM approach (also known as "EMA", Ecological Momentary Assessment) starts with the observation that participants can respond more accurately to simple questions about the here-and-now (e.g., "What language are you now using? What is the L1 of the person you are talking to?") than they can respond to broader questions such as "What percentage of the time do you use English?" Aggregating responses to simple questions asked many times is expected to yield more accurate estimates of percent L2 use than items appearing on a written LBQ. It will also provide a way to determine what percentage of L2 input is foreign-accented, and to define the overall VOT input distributions to which learners of English have been exposed.

How might the ESM technique be applied to L2 research? The study we envisage would test a large number of NS participants on the campuses of multiple American universities. (The actual number needed must be determined by a power analysis.) For inclusion, participants would need to have arrived in the United States after the age of 3 years (to exclude simultaneous bilinguals), to have lived there continuously for at least one year, to be capable of a simple conversations in English, and to report using English on a regular basis. Participants enrolled in the study will also need to have a personal smartphone, which will be used to administer the protocol.

In the first of two on-campus sessions, the NS participants will give informed consent, agree to respond to notifications that arrive via their smartphone during normal waking hours, and return for a second on-campus session roughly 2.5 months later.

Notifications will arrive via the participants' smartphone for 60 successive days. Each notification will begin with the display on the participant's smartphone of a "test code" consisting of two strings of three numbers each, for example "6 10 3 / 9 2 5" or "5 2 7 / 4 10 8". Participants will be told that the test codes serve to identify the place and time of each notification. Importantly, the second and fifth digits in the test codes will always be a number beginning with /t/ (*two* or *ten*, order counterbalanced); the remaining digits will vary randomly. VOT of /t/ in the numbers *two* and *ten* will be subsequently measured, and the entire six-number strings analyzed to identity the language background of the participants' interlocutors.

Once the participants have read the test code aloud into their smartphone they will be asked if they are currently speaking to someone. If the answer is "no", the call will end and the participants' location and their recording of the test code will be uploaded. When participants are engaged in conversation at the time of a notification, the participants will be asked to have each interlocutor(s) record the test code on the participants' smartphones. VOT in the /t/-initial numbers will be measured offline and the L1 background of the interlocutors identified using techniques similar to those employed by Flege and Munro (2004) for productions of the word *taco* by English and Spanish monolinguals.

In the second on-campus session, the participants will be paid based on the number of test codes they have successfully recorded (both their own productions as well as those of interlocutors). Additional data will also be collected in the following order: production of English words beginning with /b d g p t k/ using an instrument that constrains language mode and activation (Flege, 2018, in preparation); production of the test codes mentioned earlier; a detailed LBQ; a test of the effort expended when listening to English (McGarrigle et al., 2014); and a Spanish-based test of phonological short-term memory (see MacKay et al., 2001).

The protocol just described will yield 180 measurements of each participant's VOT production (120 via the smartphone, 60 obtained during the second on-campus session), location data, as well as VOT data and the identification of the language backgrounds of the interlocutors with whom the participants typically speak English. The data will be analyzed to determine how often each participant uses English and what percentage of their English input is Spanish-accented. Most importantly, the data base will enable investigators to construct a "VOT input distribution" for each participant. These distributions will index the full range of input received from both native speakers of English and from native speakers of Spanish, some of whom are likely to produce English /t/ with VOT values that are too short by English phonetic standards.

We expect the study to yield VOT production data similar to the data in Figure 1. We can be confident, based on previous research, that inter-subject variability will be greater among Late than Early learners. We also expect, however, that some Early learners in a large sample will differ substantially from English monolinguals, just like the Early learners tested in Puerto Rico by Flege and Eefting (1987). By hypothesis, non-nativeness in Early learners' productions, if observed, will be evident most often for Early learners who obtain a substantial amount of Spanish-accented English input. Subsequent research we carried out with some of the participants in the present study indicated that this is likely to happen when NS immigrants use English in the workplace and many fellow workers are NS Late learners.

The primary question to be addressed in the hypothetical study just outlined is: How much variance can be accounted for by a maturational factor and how much by an input factor? The maturational factor will be indexed by Age of arrival in the United States as well as self-reports by participants of their age at the time they were actually *first exposed* to spoken English on a regular basis. The input factor will be indexed primarily by the VOT input distributions mentioned earlier.

A study like the one just outlined might help resolve the seemingly unending debate between empiricists and nativists regarding the nature of L2 speech learning. Such a study might also provide important new insight into the influence of language mode and activation on VOT production. A question that might be addressed is whether VOT values obtained via participants' smartphone as they go about their normal daily lives will differ from the VOT values obtained using the same smartphones during the participants' second on-campus session. More specifically: Will VOT values decrease when participants have just been using Spanish? Or when they have just been speaking English in the presence of other NS speakers?

We hypothesize that, just as in monolingual L1 acquisition and L2 learning by "pre-Critical Period" learners, L2 learning by "post-Critical Period" Late learners is also *input driven*. If so, the Late learners' VOT production values should mirror the distributions of VOT values to which they have been exposed. It will be possible to reject a maturational account of L2 speech learning if input accounts for substantially more variance in VOT production values than does a maturational ("age") factor, especially if the "effort" and PSTM tests fail to distinguish Late learners who produced VOT in English /t/ accurately and inaccurately.

In summary, the present study revealed little difference between native Spanish Late learners who had received 0.2 vs 3.0 years of English input. The essentially null finding of the present study might be attributed either to a decreased capacity for learning new forms of speech following the closure of a critical period or to a lack of adequate L2 input. Choosing between these two explanations will require additional research that provides accurate measures of the quantity and quality of L2 input received.

Acknowledgments

This research was supported by a grant to the University of Alabama at Birmingham from the National Institute for Deafness and Other Communicative Disorders (DC00257). This article is dedicated to our former colleague Susan Guion Anderson.

References

Aslin, R., & Pisoni, D. (1980). Some developmental processes in speech perception. In G. Yeni-Komshian et al. (Eds.) *Child phonology, Volume 2, Perception* (pp. 67-96). New York: Academic press.

Bassetti, B. (2008). Orthographic input and second language phonology. In T. Piske & M. Young-Scholten (Eds.), *Input Matters in SLA*. (pp. 191-206). Bristol: Multilingual Matters.

Bassetti, B. (2017). Orthography affects second language speech: Double letters and geminate production in English. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1835-1842.

Bates, E., & Elman, J. (1996). Learning rediscovered. *Science,* 274(5294), 1849–1850.

Bruer, J. (2001). A critical and sensitive period primer. In D. Bailey, J. Bruer, F. Symons, & J. Licthman (Eds.), *Critical thinking about critical periods* (pp. 3-26). Baltimore, MD: Brookes Publishing Co.

Cowie, F. (1999). *What's within? Nativism reconsidered.* Oxford: Oxford University Press.

Csikszentmihalyi, M. & Larson, R. (1987). Validity and reliability of the experience-sampling method. *Journal of Nervous & Mental Disease*, 175, 526–537.

Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *Journal of the Acoustical Society of America*, 116(6), 3668-3678.

Dalbor, J. (1980). *Spanish pronunciation, Theory and practice*. New York: Holt, Rinehart and Winston.

DeKeyser, R. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, 22, 499-533.

DeKeyser, R., & Larson-Hall, J. (2005). What does the critical period mean? In J. Kroll & A. DeGroot (Eds.), *Handbook of bilingualism, Psycholinguistic approaches* (pp. 88.108). Oxford: Oxford University Press.

Dimitrieva, O., Llanos, F., Shultz, A., & Francis, A. (2015). Phonological status, not voice onset time, determines the acoustic realization of onset f0 as a secondary cue in Spanish and English. *Journal of Phonetics*, 40, 55-95.

Docherty, G., Watt, D., Llamas, C., Hall, D., & Nycz, J. (2011). Variation in voice onset time along the Scottish-English border. In W. Lee (Ed.) *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII): August 17-21, 2011*(pp. 591-594). Hong Kong: City University of Hong Kong.

Eguchi, S., & Hirsch, I. (1969). Development of speech sounds in children. *Acta Otolyaryngologica*, Supplement 157.

Elliott, L. (1979). Performance of children ages 9 to 17 years on a test of speech intelligibility in noise using sentence material with controlled word predictability. *Journal of the Acoustical Society of America*, 66, 651-653.

Elliott, L., Busse, L., Partridge, R., Rupert, J., & DeGraaff, R. (1986). Adult and child discrimination of CV syllables differing in Voicing Onset Time. *Child Development*, 57, 628-635.

Flege, J. E. (1987). A critical period for L2 learning to pronounce foreign languages? *Applied Linguistics*, 8, 162-177.

Flege, J. E. (1988). The production and perception of foreign language speech sounds. In H. Winitz (Ed.) *Human communication and its disorders, A review – 1988* (pp. 224-401). Norwood, N.J.: Ablex

Publishing Corporation.

Flege, J. E. (1989). Differences in inventory size affects the location but not the precision of tongue positioning in vowel production. *Language and Speech*, 32, 123-147.

Flege, J. E. (1991a). Age of learning affects the authenticity of voice onset time (VOT) in stop consonants produced in a second language. *Journal of the Acoustical Society of America*, 89, 395-411.

Flege, J. E. (1991b). Orthographic evidence for the perceptual identification of vowels in Spanish and English. *Quarterly Journal of Experimental Psychology*, 43, 701-731.

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). Methods for assessing the perception of vowels in a second language. In E. Fava & A. Mioni (Eds.), *Issues in clinical linguistics* (pp. 19-44). Padova: Uni Press.

Flege, J. E. (2008). Give input a chance! In T. Piske & Young-Scholten, M. (Eds.) *Input matters in SLA* (pp. 175-190). Bristol: Multilingual Matters.

Flege, J. E. (2018). It's input that matters most, not age. *Bilingualism: Language and Cognition*. In press.

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, 437-470.

Flege, J. E., & Eefting, W. (1986). Linguistic and developmental effects on the production and perception of stop consonants. *Phonetica*, 43, 155-171.

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, 729-740.

Flege, J. E., Frieda, E., & Nozawa, T. (1997). Amount of native-language (L1) use affects the pronunciation of an L2. *Journal of Phonetics*, 25, 169-186.

Flege, J. E., Frieda, E., Walley, A., & Randazza, L. (1998). Lexical factors and segmental accuracy in second language speech production. *Studies in Second Language Acquisition*, 20, 155-187.

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

Flege & MacKay (2004). Perceiving vowels in a second language. *Studies in second language acquisition,* 26, 1-34.

Flege, J. E., & MacKay, I. (2011). What accounts for "age" effects on overall degree of foreign accent? In M. Wrembel, M. Kul & Dziubalska-Kołaczyk, K. (Eds.), *Achievements and perspectives in the acquisition of second language speech. New Sounds 2010*, Vol. 2 (pp. 65-82). Bern: Peter Lang.

Flege, J. E., MacKay, I., & Meador, D. (1999). Native Italian speakers' production and perception of English vowels. *Journal of the Acoustical Society of America*, 106, 2973-2987.

Flege, J. E., & Munro (1994). The word unit in L2 speech production and perception. Studies in second language acquisition, 16, 381-411.

Flege, J. E., Munro, M., & MacKay, I. (1995). Factors affecting strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America,* 97, 3125-34.

Flege, J. E., Schirru, C., & MacKay, I. (2003). Interaction between the native and second language phonetic subsystems. *Speech Communication*, 40, 467-491.

Flege, J. E., & Schmidt, A. (1995). Native speakers of Spanish show rate-dependent processing of English stop consonants. *Phonetica,* 52, 90-111.

Flege, J. E., Schmidt, A., & Wharton, G. (1996). Age of learning affects rate-dependent processing of stops in a second language. *Phonetica,* 53, 143-161.

Fox, R. A., Flege, J. E., & Munro, M. (1995). The perception of English and Spanish vowels by native English and Spanish listeners: A multidimensional scaling analysis. *Journal of the Acoustical Society of America,* 97, 2540-2551.

Frieda, E., Walley, A., Flege, J. E., & Sloane, M. (1999). Adults' perception of native and nonnative vowels: Implications for the perceptual magnet effect. *Perception & Psychophysics*, 61(3), 561-577.

Garcia, P., & Froud, K. (2018). Perception of American English vowels by sequential Spanish-English bilinguals. *Bilingualism: Language and Cognition*, 2(1), 80-103.

Granena, G., & Long, M. (2012). Age of onset, length of residence, language aptitude, and ultimate attainment in three linguistic domains. *Second Language Research*, 29(3), 311-343.

Guion, S., Flege, J. E., Akahane-Yamada, R., & Pruitt, J. (2000). An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *Journal of the Acoustical Society of America*, 107(5), 2711-24.

Hazan, V., & Barrett, S. (1999). The development of phoneme categorization in children aged 6 to 12 years. In J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. Bailey (Eds.) *Proceedings of the 14th international congress of phonetics sciences* (pp. 2493-2496). Berkeley, CA: Department of Linguistics, UCLA.

Hyltenstam, K., & Abrahamsson, N. (2003). Maturational constraints in SLA. In C. Doughty & M. Long (Eds.) *Handbook of second language acquisition* (pp. 539-588). London: Blackwell.

Johnson, C. (2000). Children's phoneme identification in reverberation and noise. *Journal of Speech, Language, and Hearing Research*, 43, 129-143.

Koenig, L. (2001). Distributional characteristics of VOT in children's voiceless aspirated stops and interpretation of developmental trends. *Journal of Speech, Language, and Hearing Research*, 44, 1058-1068.

Labov, W., Ash, S., & Boberg, C. (2006). The atlas of North American English. Berlin: Mouton de Gruyter.

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *Journal of the Acoustical Society of America*, 105, 1455-1468.

Lengeris, A., & Hazan, V. (2007). Cross-language perceptual assimilation and discrimination of southern British English vowels by Greek and Japanese learners of English. In W. Barry & J. Trouvain (Eds.) *Proceedings of the 16th international congress of phonetic sciences* (pp. 1641 to 1644). Saarbruken: Saarland University.

Lenneberg, E. (1967). *Biological foundations of language*. New York: Wiley.

Long, M. (1990). Maturational constrains on language development. *Studies in Second Language Acquisition*, 12(3), 251-285.

MacKay, I., Meador, D., & Flege, J. E. (2001). The identification of English consonants by native speakers of Italian. *Phonetica*, 58, 103-125.

McGarrigle, R. Munro, K., Dawes, P. Stewart, A., Moore, D., Barry, J., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? *International Journal of Audiology*, 53(7) 433-440.

Moyer, A. (2008). In Piske, T. & Young-Scholten, M. (Eds.) *Input matters in SLA* (pp. 159-174). Bristol: Multilingual Matters.

Piske, T. & Young-Scholten, M. (2008). *Input matters in SLA*. Bristol: Multilingual Matters.

Schmidt, A., & Flege, J. E. (1995). Effects of speaking rate changes on native and nonnative production. *Phonetica*, 52, 41-54.

Schmidt, A., & Flege, J. E. (1996). Speaking rate effects on stops produced by Spanish and English monolinguals and Spanish/English bilinguals. *Phonetica*, 53, 162-179.

Sampson, G. (2005). *The 'Language Instinct' debate.* New York: Continuum.

Scholes, R. (1967). Phonemic categorization of synthetic vocalic stimuli by speakers of Japanese, Spanish, Persian and American English. *Language and Speech*, 10, 46-68.

Snodgrass, J., Levy-Berger, G., & Haydon, M. (1985). *Human experimental psychology*. Oxford: Oxford University Press.

Stölten, K., Abrahamsson, N., & Hyltenstam, K. (2014). Effects of age of learning on voice onset time: categorical perception of Swedish stops by near-native L2 speakers. *Language and Speech*, 57(4), 425-450.

Tomasello, M. (2009). *Constructing a language: A usage-based theory of language acquisition*. Boston: Harvard University Press.

Van Alphen, P, & Smits, R. (2004). Acoustical and perceptual analysis of the voicing distinction in Dutch initial plosives: the role of prevoicing. *Journal of Phonetics,* 32, 455-491.

Van Alphen, P, & McQueen, J. (2006). .The effect of voice onset time differences on lexical access in Dutch. *Journal of experimental psychology: Human perception and performance*, 32(1), 178-196.

Whiteside, S., & Dobbin, R. H. (2003). Patterns of variability in voice onset time: a developmental study of motor speech skills in humans. *Neuroscience Letters*, 347(1), 29-32.

Yeni-Komshian, G., Flege, J. E., & Liu, S. (2002). Pronunciation proficiency in the first and second languages of Korean-English bilinguals. *Bilingualism: Language and Cognition*, 3(2), 131-149.