What accounts for “age” effects on overall degree of foreign accent?

James Emil Flege and Ian R. A. MacKay
University of Alabama at Birmingham, University of Ottawa

ABSTRACT

Four hypotheses regarding “age” effects on ultimate L2 proficiency are considered here. All four are found to have some predictive power, but none is perfect. It appears that age-related effects arise from multiple factors that co-vary with the age at which L2 learning began. Of these, the amount and quality of L2 input received, and the strength of the L1 system, may be the most important long-term determinants of ultimate L2 proficiency.

KEYWORDS: L2 speech, AOA, age, foreign accent, critical period, input, individual differences

1. Introduction

That “earlier is better” for learning a second language (henceforth “L2”) has been demonstrated convincingly from the 1980s on. An “age” factor seems to be especially strong for segmental aspects of L2 speech perception and production. By now hundreds of studies have examined how particular vowels and consonants (or simply “sounds”, for short) are identified or discriminated by learners of an L2, how L2 learners actually articulate the sounds making up L2 words, and the consequences of an inaccurate phonetic rendition of L2 words by non-native speakers, for example, the difficulty experienced by listeners trying to recognize foreign-accented words.

The prevailing view of L2 speech learning, especially that of non-academics and non-specialists, can be summarized as follows. Everyone (or nearly everyone) who begins to learn an L2 after puberty is destined to forever speak it with a foreign accent (FA) whereas anyone lucky enough to begin learning an L2 before the end of a “critical period” will learn the L2 effortlessly, rapidly, and perfectly. This view is not without an element of truth. However, as will be discussed here, it is mistaken in some important respects. The aim of this contribution, therefore, is to briefly review some key findings of research on age-related effects on L2 learning and then to define an important goal for future research.

1.1. AOA in L2 research

In L2 research, “age” usually refers to the *chronological* age at which L2 learning began. Much of our work in the period 1985-2005 examined immigrants whose age of L2 learning was their age of arrival (henceforth “AOA”) in the host country. One study, for example, examined native speakers of Italian who learned English in Canada after emigrating there from Italy (Flege et al. 1995). A similar study examined Koreans who learned English in the United States (Flege et al. 1999). Both studies focussed on overall degree of FA in the pronunciation of English sentences. With few exceptions, the participants' contact with English was immediate and plentiful because most were compelled to use English for everyday communication.

In immersion studies like these, AOA marks the age of participants' first substantial contact with the L2 as it is spoken by native speakers. In the two studies just cited, a few participants who arrived as adolescents or young adults had studied English briefly in school before emigrating. In our view, however, such exposure to the L2 — usually in the form of the foreign-accented renditions of a non-native teacher — probably did not exert an important or lasting effect. Individuals who arrived in the first two years of life were excluded from both studies to ensure that the research focused on sequential rather than simultaneous bilingualism.

One desirable characteristic of AOA is that it can be reported accurately by participants. Immigration is a very important, often traumatic, event in the life of an individual. Virtually without exception, individuals who participated in several of our studies in Ottawa over a 10-year period reported the same year and month of arrival in Canada, and usually the same day of arrival! Importantly, participants differing in AOA can usually be recruited given a sufficient amount of patience, at least within the confines imposed by inherent characteristics of the local immigration population. For example, we discovered empirically that it was fairly easy to locate individuals with AOAs of 2 to 8 years and with AOAs of 18 to 23 years. However, it proved impossible to find Italian immigrants in Ottawa having AOAs greater than 35 years.

AOA might seem simple inasmuch as it indexes a participant's age at a particular moment in time, but it is in fact a complex variable. Researchers choose to use AOA in their research, either explicitly or implicitly, because of AOA's presumed association with other, underlying variables which are hypothesized to directly impact L2 learning. These "underlying" variables include participants' overall state of neural maturation when L2 learning began, the state of development of native language (L1) phonetic categories, and the kind and/or amount of L2 input typically received. Given that multiple simple variables are associated with AOA, and given that many of these simple variables are inter-correlated, AOA must be regarded as a "macro-variable".

Researchers have frequently computed correlations between AOA and various outcome measures of interest. One example is the seminal study by Johnson and Newport (1989) These authors examined knowledge of 12 important grammatical structures of English as represented by grammatical and ungrammatical pairs of 138 English sentences. The participants were 46 Chinese and Korean speakers, all students or professors at an American university. To be included, participants needed to have lived in the U.S. for at least 3 years. The experimental task was to label each sentence as "grammatical" or "ungrammatical". The dependent variable was the percentage of correct responses. A strong correlation between AOA and the percent correct scores was obtained: the later the arrival, the lower the scores. Further, the AOA-performance correlation was found to be significant for participants who had arrived in the U.S. by the age of 15 years but non-significant for those with AOAs of 17 or greater. This finding led the authors to conclude that a "maturational constraint" exists for L2 learning.

Investigators often derive variables in addition to AOA from a language background questionnaire administered to their participants. There has naturally been interest in determining if AOA is a stronger predictor of L2 performance than are other variables but, alas, this objective is usually impeded by the existence of multi-collinearity between the multiple questionnaire variables. For example, Flege et al. (1995) obtained significant correlations between AOA and degree of FA (the earlier the arrival, the milder the FAs), between years of residence in Canada and FA (the longer, the milder), and between AOA and years of residence (the earlier, the longer). The AOA-FA correlation was much stronger than was the AOA-years of residence correlation. Thus if AOA were used as a surrogate for "degree of neural plasticity" at the time L2 learning began, and if years of residence were used as a surrogate for "overall amount of L2 input received", a researcher might be tempted to conclude that brain plasticity is a more important predictor of ultimate accuracy in pronouncing an L2 than is input.

However, as discussed by Flege (2009), such a conclusion would be unwarranted. While it is possibly true that the earlier arrivals tended to have more "plastic" brains than the later arrivals did when L2 learning began, the earlier arrivals were *also* likely to have received more input from L2 native speakers than the later arrivals had. The two predictor variables were correlated, and thus not independent of one another. Moreover, given that the two predictors were based on different metrics, a head-on comparison is fraught with peril. Is AOA a better or worse index of "neural plasticity" than years of residence was an index of "total amount of L2 "input"? Can we be sure that the ranges of values for the two variables were functionally equivalent? Given that it is difficult if not impossible to answer questions of this kind, conclusions must be reached very cautiously.

Investigators who use a correlation approach have tended to impose few restrictions on participant selection because of the need to test as many participants as possible. Indeed, not all published studies have required that all participants speak the same native language (L1), or have had some minimum demonstrable amount of experience actually using the target L2. And few studies have placed upper limits on participants' chronological age or everyday use of the L1. Another frequent methodological oversight is a failure to tell readers how the research participants were recruited. This is unfortunate inasmuch as recruitment procedures may lead to subtle biases that, in turn, influence the outcome of a study.

Why is this? Participants in L2 research are rarely if ever sampled randomly. They are usually recruited through a social network of a researcher's friends and acquaintances and/or through the participants' affiliation with a school or university. For example, Flege et al. (1995) identified potential native Italian participants living in Ottawa, Ontario through their connection — currently or in the past — with a particular Roman Catholic church that has for a very long time been a focal point of the Italian community in Ottawa. Contact with potential participants was made by an Italian-Canadian woman who was herself a long-time member of the church. Flege et al. (1999), on the other hand, recruited Korean immigrants primarily on the campus of the University of Maryland or from the congregations of nearby Korean churches frequented by members of the university community. Korean-American students at the university served as recruiters. These differing procedures may have contributed to between-study differences in years of formal education in English-medium schools and universities. More Korean than Italian participants completed high school in the host country, and far more Korean than Italian participants had obtained a college degree in the host country.

Other biases might be introduced by inherent characteristics of the immigrant population under investigation. For example, few Italians immigrated to Canada after 1970 whereas few Koreans immigrated to the United States *prior* to that year. Not surprisingly, the Italians tested by Flege et al. (1995) had lived in Canada far longer (*mean* = 32 years) than the Koreans tested by Flege et al. (1999) had lived in the United Sates (*mean* = 15 years).

Other studies have used ANOVA techniques to examine the effect of AOA. These studies are often said to employ a "retrospective developmental" design because they compare groups of adults who differ according to the age when, in the past, they began learning an L2. For example, Flege (1991) recruited groups of native Spanish adults who had arrived in the United States at average ages of 2 vs. 20 years. The two groups were referred to as relatively "early" and "late" learners of English L2. The study focussed on one specific aspect of speech production, namely voice-onset time (VOT) in word-initial stop consonants. The early learners were found to produce English /t/ with mean VOT values that were very similar to those observed for English monolinguals whereas the late learners' VOT values were intermediate to the mean values obtained for English and Spanish monolinguals

The Flege et al. (1995) study cited earlier made use of a more elaborate version of the same basic design. The authors recruited native Italian (NI) immigrants whose AOAs fell within of one of 10 abutting ranges, 1.9 to 4.1 years (*mean* = 3.1) for the earliest-arriving group and 20.2 to 23.2 years (*mean* = 21.5) for the latest-arriving group. All 10 NI groups and also the native English (NE) comparison group consisted of 24 participants. The dependent variable was overall degree of perceived FA in English sentences. One aim of this study, as well as similar study examining Korean immigrants to the United States (Flege et al. 1999), was to identify the AOA of the first group to differ significantly from the NE comparison group. This study was the first to convincingly demonstrate that "pre-critical period" L2 learners, that is, individuals who had begun learning the L2 prior to the age of 12 years, continued to speak the L2 with a detectable FA even after years of L2 immersion.

2. A new FA experiment

AOA effects on L2 performance are generally robust, but the best explanation for such effects remains uncertain, even controversial. With this in mind, we carried out a new FA study. Its aim was to test predictions generated by four hypotheses regarding the underlying bases of the AOA effect on degree of FA.

2.1. Participants

Some of the participants for this study were drawn from a previously published study that focussed on chronological age and sentence duration (MacKay et al., 2006). Others were recruited specifically for this study, the aim of which was to extend the range AOA values of the native Italian (NI) participants tested.

The decision was made to exclude any NI participant who arrived in Canada before age 7. This was done to avoid comparing individuals who learned English as an L2 (sequential bilinguals) to those who had two "first" languages (simultaneous bilinguals). Another motive for this decision was the possibility that AOA difference between individuals who arrived in Canada prior to school age might not be meaningful. Our informal conversations with members of the Italian community in Ottawa suggested that young NI children oftentimes had relatively little exposure to English before going to school.

The mean AOA of the latest-arriving group tested by Flege et al. (1995) was 21.5 years. Here we managed to recruit a group of NI participants who arrived somewhat later in life. Specifically, we recruited 18 NI participants each whose AOAs ranged from 7 to 13 years (*mean* = 10.4 years), 17 to 19 years (*mean* = 18.1 years), and 23 to 35 years (*mean* = 25.7 years). These groups will be referred to using the labels "AOA-10", "AOA-18" and "AOA-26".

Table 1. Characteristics (means, SDs) of the three native Italian groups. The effect of Group was significant in all instances ($p < .01$); between-group differences were evaluated using a Tukey test ($p < 0.05$).

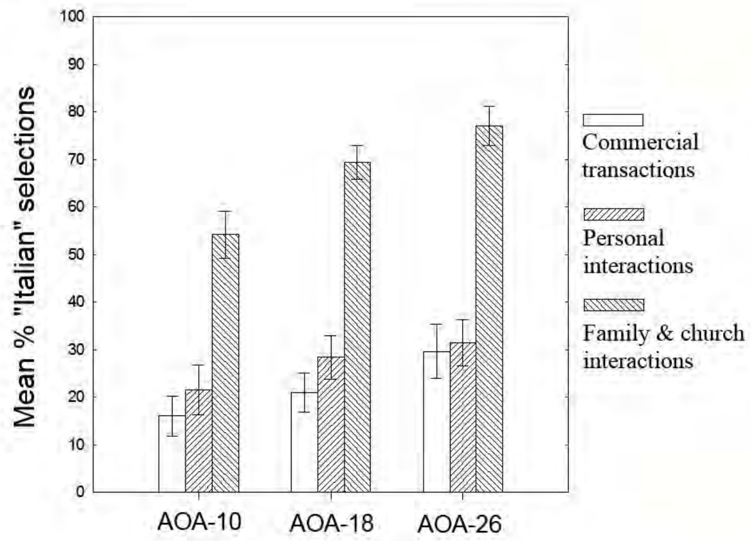
	AOA-10	AOA-18	AOA-26	Tukey test
Residence in Canada (years)	42 (5)	37 (7)	33 (8)	10 > 18,26
Self-reported use of English (%)	71 (14)	53 (15)	47 (19)	10 > 18,26
Education in Canada (years)	9.0 (3.8)	0.4 (1.3)	0.3 (0.6)	10 > 18,26
Self-reported use of Italian (%)	27 (14)	47 (15)	52 (20)	10 < 18,26
Education in Italy (years)	4.1 (2.4)	7.9 (2.4)	7.6 (2.9)	10 < 18,26
Italian use in context (%)	30 (17)	39 (15)	46 (19)	10 < 26
Self-ratings (7-point scale)				
Speak and understand English	6.5 (0.6)	5.8 (0.9)	5.4 (1.0)	10 > 18,26
Pronunciation of English	6.0(1.0)	4.8(1.2)	4.4 (1.1)	10 > 18,26
Speak and understand Italian	5.5 (1.0)	6.4 (0.8)	6.8 (0.3)	10 < 18,26
Pronunciation of Italian	5.3 (1.1)	6.4 (0.8)	6.8 (0.5)	10 < 18,26

All 54 NI participants and all 18 members of the NE comparison group were recruited in Ottawa, Ontario and passed a pure-tone hearing screening at octave frequencies between 500-4000 Hz (re: 35 dB HL). The NE group had much the same gender makeup (7 males) as did the three NI groups (7 to 9 males per group). The NE participants had all been born and raised in Canada and did not speak any language other than English well or often. The NI participants did not speak English prior to immigrating to Canada, and had been living there for 13 to 53 years when tested. The mean chronological ages of the groups ranged from 53 years (NE) to 59 years (AOA-26). This was the only between-group difference in age to reach significance ($p < 0.05$ by Tukey's test). However, given evidence that such small differences in chronological age do not affect FA (MacKay et al., 2006), this variable will not be discussed further.

Table 1 summarizes the NI participants' responses to a language background questionnaire. The effect of Group (3 levels) was significant for all variables ($p < .01$). Tukey tests indicated that, in all but one instance, the AOA-18 and AOA-26 groups differed from the AOA-10 group but did not differ from one another ($p < .05$). Compared to the two later-arriving groups, the participants in AOA-10 had arrived earlier in Canada, had lived longer in Canada, had attended Canadian schools longer, and reported using English more often. Also, they reported being better able to speak, understand, and pronounce English than the later-arriving participants in AOA-18 and AOA-26 but, conversely, reported less ability to speak, understand, and pronounce Italian.

The only questionnaire variable for which the AOA-10 and AOA-18 groups did not differ significantly was the one called "Italian use in context". This novel variable, like "percentage Italian use", was intended to index frequency of L1 use. It was computed by asking participants to indicate whether they would be more likely to use English or Italian in contexts representing a wide range of everyday activity. (Seven contexts had to be eliminated because five or more participants indicated that they fell outside their normal range of activity.) Surprisingly, the percentage of contexts in which Italian was chosen as the more likely language to be used showed only a moderate correlation with overall estimates of percentage Italian use , $r(52) = 0.45$, $p < 0.01$.

Figure 1. Mean percentage of times that native Italian participants differing in age of arrival (AOA) in Canada (10, 18, or 26 years) indicated being more likely to use Italian than English in various everyday situations, or "contexts" (see text). The error bars bracket +/- 1 SE.



Three sub-scores were, therefore, computed to further explore the Italian use in context variable. One set (n = 9) consisted of items that focussed on social interactions of a commercial nature such as "making travel arrangements" (at a time when travel agents were generally consulted in their offices) or "buying groceries". The second set (n = 8) included items pertaining to interactions with family members or fellow church members. The final set (n = 9) involved all other social interaction, such as "talking to a neighbour" or "visiting a close friend". As shown in Fig. 1, the NI participants reported being more likely to use Italian than English when interacting with family/church members (*mean* = 67% of contexts) than when carrying out a commercial transaction (*mean* = 22%) or when involved in other personal interactions (*mean* = 27%).

ANOVAs revealed a significant effect of Group (3 levels) for the family/church context, $F(2,51) = 7.6$, $p < .01$, but not for the commercial context, $F(2,51) = 2.1$, $p > 0.10$, or the "other" context, $F(2,51) = 1.1$, $p > .10$. A Tukey test examining the family/church scores yielded the same pattern of differences obtained for the other questionnaire variables: AOA-18 and AOA-26 both differed significantly from AOA-10 but did not differ from one another ($p < .05$). The difference between the sub-scores probably reflects the linguistic make-up of Ottawa, where Italian-speaking immigrants represent a very small percentage of the overall population (< 1%). We suspect that our NI participants had much more opportunity to use Italian when at home or at church than when conducting various commercial transactions or encountering other citizens "on the street".

One might infer from the questionnaire data that participants in AOA-18 and AOA-26 had less overall exposure to English than those in AOA-10 during their first years in Canada, but more exposure to Italian-accented English. Participants in AOA-10 all began attending English-medium school soon after arriving in Canada. While attending school they had many opportunities — in fact, a strong need — to interact with native speakers of English, and they were likely to have formed life-long friendships with some of their schoolmates.

Given that our NI participants immigrated to Canada in order to better their economic condition, it seems reasonable to think that all of the males in AOA-18 and AOA-26 entered the work force soon after arriving there. We infer that they sought work through social networks in the Italian-speaking community, and that many found jobs permitting them to continue using Italian. Female members of these groups — especially those who had young children — were probably less likely to work outside the home in the years immediately following immigration. Perhaps their husbands' Italian-accented English provided their first model of English pronunciation.

The inferences just drawn from the questionnaire variables are consistent with two findings reported by Flege et al. (1995). These authors found that for individuals who arrived in Canada at the age of 10 years, females had significantly milder FAs than did males, thereby demonstrating the expected "female" advantage in L2 learning. However, for individuals with an AOA of 21 years, FAs were significantly milder for males than females. Second, males' self-reported use of English in the workplace accounted for 9% of the variance in FA ratings whereas this variable was not found to be a significant predictor of females' pronunciation of English.

2.2. Method

As in previous research (e.g., Piske et al., 2001), we used a delayed repetition technique to elicit the production of English sentences. Subsequent auditory inspection of the sentence corpus revealed the presence of a number of disfluencies and pauses, not to mention failures by some of the late-arriving NI participants to repeat the target sentences word for word. Also, some recordings were marred by peak clipping (for further detail, see MacKay et al., 2006).

The decision was made, therefore, to focus on three portions of sentences that were free of any such artefacts in the productions of all participants. The phrases examined were the underlined portions of the following sentences: *We were neighbours for thirty years*; *Billy is Patty's husband*; and *We'll have pork chops and potatoes*. Three replicate productions of each phrase by each participant were edited out of the original sentences and normalized for peak intensity, yielding 648 phrases for analysis (72 participants x 3 phrases x 3 tokens).

The test phrases were presented in three separate randomizations in sessions held on three separate days to each of 18 native-English speaking listeners from Ottawa. The listeners used a scale ranging from 1 ("very strong foreign accent") to 9 ("no foreign accent") to rate each phrase. No feedback or training was given in the practice block administered before testing began on the first day. The listeners were urged to use the entire scale and were told to guess if uncertain. Test phrases could be replayed, but ratings could not be changed once given.

2.3. Results and discussion

As shown in Fig. 2, the FA ratings decreased — indicating increasingly strong FAs — in the following order: NE > AOA-10 > AOA-18 > AOA-26. To test for between-group differences, we calculated a single FA rating for each of the 72 participants. Each mean was based on 486 ratings (18 listeners x 3 phrases x 3 tokens x 3 replicate judgments). The significant effect of Group on the ratings, $F(3,68) = 120.8$, $p < .0001$, was explored by a Tukey test. The post-hoc test revealed that all pair-wise differences between the four groups were significant ($p < .001$). In subsequent sections we discuss how these findings impact each of four hypotheses regarding the basis of AOA on L2 speech learning.

2.3.1. The maturational constraint hypothesis (H1)

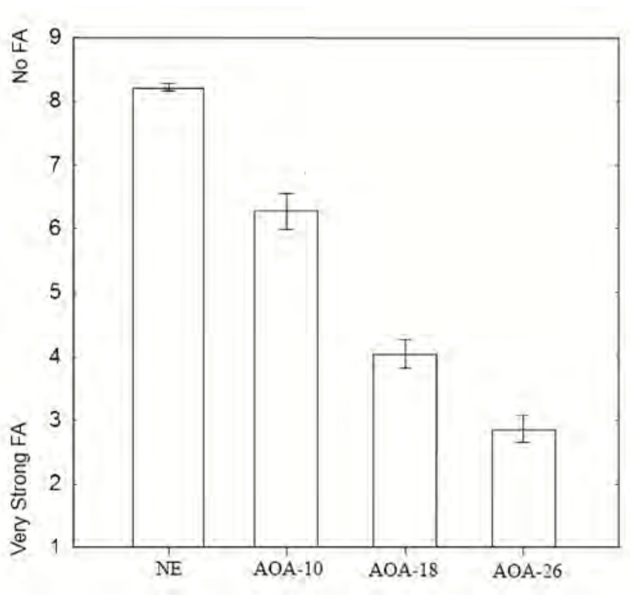
H1 states that endogenous changes that occur as humans mature make certain mechanisms used in language acquisition less effective. H1 is derived from research focussing on morphosyntax. Like Johnson and Newport (1989), DeKeyser (2000: 518-519) observed strong effects of AOA on grammaticality judgment test scores. He concluded that such effects arose as the consequence of a "severe decline *[in]* ability to induce abstract patterns implicitly" which is an "inevitable consequence of fairly general aspects of neurological maturation". Following Johnson and Newport (1989), DeKeyser fixed the end of a critical period at about 16 years. Others (e.g., Scovel, 2000), however, have suggested a somewhat earlier end of a critical period for phonological acquisition.

H1 correctly predicts the difference in FA between AOA-10 and AOA-18. The relatively stronger FAs of AOA-18 than AOA-10 might be due to the closing of a critical period between the ages of 10 and 18 years. However, H1 does not predict the difference between NE and AOA-10, which was expected from previous research (e.g., Flege et al., 1995; Flege et al., 1999).

The failure to predict the presence of FA for AOA-10 (i.e., to predict a NE vs. AOA-10 difference) is notable inasmuch as the presence of FA in “pre-critical period” learners is by now well established. Although most work showing this effect has examined adults differing in AOA, Flege et al. (2005) observed the same effect in children. Korean children’s sentences received significantly lower FA ratings than those of age-matched NE children. In fact, a detectable FA was evident even for Korean children with an average AOA of 6 years and an average chronological age of 11, that is, following an average immersion of 5 years.

Another problem for H1 is that it generates a prediction that has been falsified in several studies. If AOA effects are due to the passing of a critical period, one would expect L2 performance to decline as participants’ AOAs near the end of a critical period. However, performance should remain stable thereafter because everyone *beyond* the critical period should show equally the ill effects of having passed the critical period.

Figure 2. Mean ratings of foreign accent (FA) in English phrases spoken by native English (NE) speakers and Italian immigrants differing in age of arrival in Canada (mean AOAs of 10, 18 or 26 years). The error bars bracket +/− 1 SE.



A re-analysis of FA ratings obtained for Italian and Korean immigrants by Flege et al. (1995) and Flege et al. (1999) showed, however, that FAs continue to grow stronger beyond the critical period. For both studies, AOA-FA correlations were computed for all 240 participants, then for just those participants with AOAs greater than 12, and finally just those with AOAs greater than 15. As the number of the participants in the samples decreased, the AOA-FA correlations necessarily weakened but, importantly, they remained significant ($p < .01$) even in the last sample which included only "post-critical period" learners.

The doubts raised here concerning H1 prompts us to return to the morphosyntax data of Johnson and Newport (1989) mentioned earlier. The strongest support obtained by these authors for H1 was a significant correlation that was found to exist between AOA and the test scores obtained for participants having an AOA of less than 16 years, but not for those with AOAs of 16+. Crucially, the existence of a critical period at age 16 was assumed, not demonstrated. However, Bialystok and Hakuta (1994) showed that when the "cut point" dividing pre- from post-critical period learners was arbitrarily shifted from 16 to 20 years, a moderately strong correlation ($r = -.49$) between AOA and the morphosyntax scores was obtained for post-critical period learners.

Even more strikingly, the results of Flege et al. (1999) suggested that the effect of AOA on grammaticality judgment test scores may disappear when factors confounded with AOA have been controlled. These authors used a test modelled on that of Johnson and Newport (1989). As in the earlier study, test scores showed a strong correlation with AOA. But, as is usually the case in L2 research, AOA was correlated with other variables that might have influenced the test scores, including years of education in English-medium schools, years of residence in the United States, and self-reported use of English.

Flege et al. (1999) used a subgroup matching procedure to unconfound other potentially important variables. To begin, two subgroups of 20 participants each having mean AOAs of 10 and 17 years were randomly selected from the original sample of 240. The "early" sub-group obtained much higher scores than did the "late" sub-group. This is hardly surprising given that the sub-groups, having been randomly selected from the larger sample, necessarily yielded much the same effects that were seen in the original sample of 240. Next, two additional sub-groups of 20 participants each were selected. These new sub-groups had the same mean AOA values of the earlier ones (viz., 10 and 17 years). However, participants for the new sub-groups were selected in such a way as to match them for years of formal education in the United States (*mean* 11 years for both). The matching process, in addition to equalizing the new early and late sub-groups for education, also eliminated significant early-late differences in years of residence and self-reported English use.

The results obtained for the matched sub-groups were quite different from those obtained for the randomly selected early and late sub-groups. The small numerical difference in morphosyntax scores obtained for the *matched* early and late learners did not even approach statistical significance. This leads one to wonder if the effect attributed by Johnson and Newport (1989) to differences in cognitive maturation for participants differing in AOA was, in fact, due to *other* variables associated with AOA such as amount of formal education in the host country or variation in L2 input.

2.3.2. The cognitive development hypothesis (H2)

H2 states that L2 learning becomes gradually less effective across the entire life span because the cognitive abilities needed for speech and language learning diminish slowly across the life span. H2 is derived from a study by Hakuta et al. (2003). These authors derived estimates of English proficiency from several responses given to the United Sates Census by large numbers of Chinese- and Spanish-speaking immigrants. Self-estimates of English-language proficiency by members of both immigrant populations were found to decrease continuously from 10 to 60 years, not just until the middle of the second decade of life as one might expect if fully effective learning can only occur prior to the closing of a critical period.

H2 correctly predicts the increasingly strong FAs seen in Fig. 2, that is, the differences observed between all three NI groups. But for H2 to really explain these differences, it will be necessary to identify a specific cognitive change(s) responsible for such differences, which has never been done. Perhaps more importantly, H2 is unlikely, in whatever form it might eventually take, to explain the difference between NE and AOA-10. It is *a priori* difficult, for us at least, to imagine 10-year-olds having undergone a cognitive change that could reduce their ability to learn an L2 whilst the L1 is still developing.

2.3.3. Changes in L1-L2 interactions (H3)

H3 states that as the L1 system develops, the effect of cross-language phonetic interference becomes stronger because of differences in how the L1 and L2 sound systems interact. This hypothesis is derived from Flege's Speech Learning Model (e.g., Flege 2002, 2003, 2007, 2009).

H3 assumes that L1 categories develop slowly into early adolescence as shown, for example, by studies examining the recognition of L1 words in noise. According to H3, as L1 categories develop they become stronger "attractors" for sounds later encountered in an L2. This has the effect of making learners less likely to create new categories for L2 sounds that are similar but non-identical to corresponding L1 sounds, even when the cross-language phonetic differences are audible.

H3 generates predictions that can be readily tested. It predicts, for example, that as native Spanish speakers get older, their ratings of the perceived phonetic dissimilarity between syllable-initial tokens of /t/ in English ([t^h]) and in Spanish ([t]) will decrease, and their correct detections of stops with English-like VOT values that have been inserted into Spanish speech samples will decrease. Crucially, H3 predicts that as an indirect consequence of L1 phonetic category development, more Spanish early than late learners will establish a new phonetic category for English /t/. This, in turn, leads to the expectation that more early than late learners will realize English /t/ with English-like VOT values.

A less obvious prediction — also derived from the SLM — regards L2 sounds for which a new category has not been formed. The Spanish sound closest to English /t/ ([t^h]) is Spanish /t/ ([t]). Although acoustic phonetic differences between this L1-L2 pair are audible, some native Spanish learners of English (by hypothesis, more late than early learners) will not attend to these audible differences and will, therefore, not begin creating a new phonetic category. SLM hypothesizes that, in such instances, learners will develop a "composite" L1-L2 phonetic category sharing properties of the L1 and L2 sounds that have been perceptually "equated". This leads to the prediction that late Spanish learners of English will tend to produce English /t/ with VOT values that are too short for English while at the same time producing Spanish /t/ with VOT values that are *too long*.

There is some support for this prediction. Flege (1991) found that early Spanish learners of English produced English /t/ with native-like values whereas late learners produced values intermediate to those observed for Spanish and English monolinguals. H3 generates the same predictions for French as for Spanish. In fact, Flege (1987) observed compromise VOT values for late French-English bilinguals; he also found that both French learners of English and American learners of French realized their L1 /t/ with values that had moved in the direction of typical the L2.

H3 might be used to account for the difference between NE and AOA-10. Depending on the time-course of L1 phonetic category development, which has never been adequately defined, H3 might also account for the difference between AOA-10 and AOA-18. However, this hypothesis can clearly not account for the difference between AOA-18 and AOA-26 because it is likely that L1 categories are fully developed in monolinguals by the age of 18 years.

A serious problem for H3, moreover, is that little research has been conducted thus far to determine if the perceived phonetic dissimilarity of pairs of L1 and L2 sounds actually *does* decrease as L1 categories develop, and whether such changes — if they do occur — predict accuracy of L2 segmental production and/or perception (but see Baker et al. 2002).

2.3.4. Input differences between early and late learners (H4)

H4 predicts that L2 performance depends on amount of L2 input, which varies as a function of both years of L2 use and frequency of daily use. (Should this hypothesis be developed further in the future, it may be necessary to give relatively greater "weight" to input received in initial phases of L2 learning.)

H4 is derived from research examining Italian immigrants in Canada (Flege and MacKay 2004; Piske et al. 2001, 2002; MacKay et al. 2001). Groups differing in AOA (*means* 8 vs. 20 years) were subdivided according to self-reported frequency of Italian use (*means* of 7% and 10% vs. 43% and 53%). For both early and late learners, a frequent use of Italian (and thus an infrequent use of English) was associated with a significantly poorer identification of word-final English consonants, a poorer discrimination of English vowels, a less accurate production of English vowels, and with relatively stronger FAs.

As discussed previously (see Table 1), it is likely that members of AOA-10 received more L2 input than did members of the other two groups. That being the case, H4 correctly predicts the difference between AOA-10, on the one hand, and AOA-18 and AOA-26, on the other hand, but it fails to predict the difference between these two latter groups.

3. Creating a predictive model

All four hypotheses considered above were able to predict some of the between-group differences in the experiment reported here, but no one hypothesis predicted that entire pattern of results obtained. The only certain conclusion that we can draw is that, ultimately, it will be necessary to develop a multi-factor model to explain why early learners generally outperform late learners.

As already mentioned, the primary problem faced by anyone who wants to develop such a model is that many of the variables likely to impact L2 performance are inter-related. As described by Flege (2009), one solution to the problem of multi-collinearity is to submit language background questionnaire variables to a principle components analysis (PCA). The PCA will yield a smaller set of "underlying" factors that, being uncorrelated with one another, can then be used as predictor variables in a multiple regression analysis that has as its aim to account for as much variance as possible in the dependent variable of interest.

Flege (2009) used this approach to re-analyze the FA ratings obtained for 240 NI immigrants to Canada (Flege et al., 1995). Six variables were derived from the questionnaire administered to the participants. The first three were based on responses to a single item: "AOA", "Years of residence in Canada", and "Chronological age". The remaining three variables were derived from responses to multiple questionnaire items: "Frequency of L2 use" was an average of self-rated use of English in the home, in social settings, and overall in the 5 years preceding the study. "L1 proficiency" was computed from items dealing with the ability to communicate in Italian via telephone, to tell jokes in Italian, to pronounce Italian, and to remember how Italian words are pronounced. Finally, a variable labelled "Motivation" was based on items dealing with the importance of L2 pronunciation and how much attention the participants paid to their own pronunciation of the L2.

All six variables except Motivation showed a significant simple correlation with degree of FA. A good pronunciation of English was associated with a relatively early arrival in Canada, a relatively young age at test, a lengthy residence in Canada, frequent use of English, and relatively poor self-estimated L1 proficiency. As expected, the variables derived from the questionnaire were inter-correlated. For example, a frequent use of English was associated with a relatively poor self-reported proficiency in the L1 which, in turn, was associated with a relatively lengthy residence in Canada and an infrequent use of Italian.

A PCA analysis was carried out to examine the six variables just described. It identified two underlying factors. Factor 1 was had high loadings (i.e., was defined by) by AOA, L1 proficiency, and Frequency of L2 use. Factor 2 had high loadings on Chronological age and Years of residence. The two factor scores accounted for 72% of the variance in FA ratings in a multiple regression analysis.

Far more variance was accounted for by Factor 1 than by Factor 2 (65% vs. 7%). Given evidence that chronological age does not affect degree of FA (MacKay et al. 2006) we might interpret Factor 2 as an index of years of residence in Canada, which is to say, as an index of L2 input. It is more difficult to interpret Factor 1, which was defined by AOA, L1 proficiency, and a Frequency of L2 use.

Given that AOA is a macro-variable with no real predictive power in itself, Flege (2009) decided to remove AOA from a second PCA analyses, which again identified two underlying factors. Factor 1 was defined by L1 proficiency and Frequency of L2 use, and Factor 2 was again defined by Chronological age and Years of residence in Canada. The two factors accounted for 57.6% of the variance in the FA ratings in a multiple regression analysis. Once again, Factor 1 accounted for far more variance (50.2%) that did Factor 2 (7.4%).

What is the best way to interpret this two-factor model? As already stated, F2 is best interpreted as an index of L2 input, that is to say, years of L2 use. The longer the NI immigrants had resided in Canada, the more years they had used English to communicate. F2 was based on L1 proficiency and frequency of L2 use. "Frequency of L2 use" might be regarded as a second, complementary index of L2 input in the sense that overall input in an L2 depends both on how long the L2 has been used as well as how frequently. If this is so, then we might conclude that the accuracy of one's eventual pronunciation of an L2 will depend on two general factors: how much L2 input has been received (the more, the better) and how strong or active the L1 remains (the stronger/more active the L1, the worse the L2 pronunciation).

Flege (2009) obtained results for a sample of 240 Korean immigrants to the United States that were very similar to those just presented for the Italians who had immigrated to Canada. This replication reinforces the conclusions just reached but they will, nevertheless, need to be confirmed by additional research. But even if the conclusion presented here is validated by additional research we are still left with a daunting question: What accounts for the remaining variance?

We can only speculate about how future research might go about developing a more adequate multifactor model to predict FA in experienced speakers of an L2, not to mention predicting other specific aspects of L2 acquisition. We doubt that measures of motivation will add much because, in our view, motivational differences are closely tied to differences in input. However, it seems likely to us that developing better estimates of L2 input — or even obtaining actual *measurements* of L2 input — will increase the variance accounted for (see Flege, 2009, for one suggestion).

The notion that a decrease in neural plasticity at around puberty results in stronger FAs has been popular for decades now and so warrants additional attention. If someone can develop a measure of "neural plasticity" with which to assess immigrants at the time they begin learning an L2, perhaps this new measure of plasticity will be found to account for additional variance not already accounted for by the variables described earlier.

As discussed earlier, the Speech Learning Model (e.g., Flege, 2003, 2007) claims that as L1 categories develop, they become stronger "attractors" for L2 sounds, thereby lessening the likelihood of new categories being established for the L2 sounds. The absence of categories for certain L2 sounds, in turn, is thought to limit accuracy of perception and production. If it were possible to devise a measure of the overall state of development of L1 phonetic categories when L2 learning begins, this new measure might account for additional variance in L2 speech performance.

Finally, we urge researchers working on this topic to evaluate more closely the role of individual differences in working memory. Of special interest is a form of memory known as phonological short-term memory (PSTM).

MacKay et al. (2001) tested the hypothesis that PSTM would predict NI immigrants' identification of word-final English consonants. PSTM was evaluated by having the participants repeat sets of Italian-like non-words which varied in number of syllables but all had penultimate stress. The non-word stimuli were created by splicing together 2 to 5 syllables edited from real Italian words. Of special interest to the present discussion, PSTM accounted for 15% of the variance in consonant identification scores after variation in AOA, years of residence, L1 use and L1 proficiency had been controlled for statistically.

4. Conclusions

AOA shows a strong correlation with many measures of L2 performance, but it is a macro-variable with no real predictive value in itself. Instead, the importance of AOA lies in its' association with several other, underlying variables that may impact L2 learning by immigrants who learn the L2 through immersion in a predominantly L2-speaking community. Importantly, no one variable will ever be able to account for the full range of L2 performance that is typically seen.

An important aim of future L2 speech research should be to create a multi-factor model that accounts for as much variance as possible in key measures of L2 performance in highly experienced speakers of an L2. Based on the research reviewed here, we conclude that the two most prominent variables in such a model are likely to be amount of L2 input and strength/activation of the L1 system. We speculate that individual differences in phonological short-term memory may also account for an important additional amount of variance.

REFERENCES

Baker, W., Trofimovich, P., Mack, M., Flege, J. 2002. The effect of perceived phonetic similarity on non-native sound learning by children and adults. In: Skarabela, B., Fish, S., Do, A. (eds.) Proceedings of the 26th annual Boston University conference on language development, Somerville: Cascadilla Press, 36–47.

Bialystok, E., Hakuta, K. 1994. *In other words: The science and psychology of second language acquisition*. New York: HarperCollins.

DeKeyser, R. 2000. The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition* 22. 499–534.

Flege, J. 1987. The production of „new" and „similar" phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics* 15. 47–65

Flege, J. 1991. 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. 2002. Interactions between the native and second–language phonetic systems. In: Burmeister, P. et al. (eds.). *An integrated view of language development: Papers in honor of Henning Wode*. Trier: Wissenschaftlicher. 217–244.

Flege, J. 2003. Assessing constraints on second–language segmental production and perception. In Meyer, N. Schiller (eds.), *Phonetics and phonology in language comprehension and production*. Berlin: Mouton. 319–355.

Flege, J. 2007. Language contact in bilingualism: phonetic system interactions. In: Cole, J., Hualde, J. (eds.). Laboratory phonology 9. Berlin: Mouton de Gruyter, 353–382.

Flege, J. 2009. Give input a chance! In: Piske, T., Young-Scholten, M. (eds.). *Input matters in SLA. Bristol: Multilingual Matters*. 175–190.

Flege, J., Birdsong, D., Bialystok, E., Mack, M., Sung, H., Tsukada, K. 2005. Degree of foreign accent in English sentences produced by Korean children and adults. *Journal of Phonetics* 33. 263–290.

Flege, J., MacKay, I. 2004. Perceiving vowels in a second language. *Studies in second language acquisition* 26. 1–34.

Flege, J., Munro, M., MacKay, I. 1995. Factors affecting degree of perceived foreign accent in a second language. *Journal of the Acoustical Society of America* 97. 3125–3134.

Flege, J., Yeni–Komshian, G., Liu, S. 1999. Age constraints on second language learning. *Journal of Memory and Language* 41. 78–104.

Hakuta, K.; Bialystok, E., Wiley, E. 2003. Critical evidence: A test of the critical–period hypothesis for second-language acquisition. *Psychological Sciences* 14. 31–38.

Johnson, J., Newport, E. 1989. Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology* 21. 60–99.

MacKay, I., Flege, J., Imai, S. 2006. Evaluating the effect of chronological age and sentence duration on degree of perceived foreign accent. *Applied Psycholinguistics 27*, 157–183.

MacKay, I., Meador, D., Flege, J. 2001. The identification of English consonants by native speakers of Italian. *Phonetica* 58. 103–125.

Piske, T., Flege, J., MacKay, I., Meador, D. 2002. The production of English vowels by fluent early and late Italian–English bilinguals. *Phonetica* 59. 49–71.

Piske, T., MacKay, I., Flege, J. 2001. Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics* 29. 191–215.

Scovel, T. 2000. A critical review of critical period research. *Annual Review of Applied Linguistics* 20. 213–223.