Second Language Acquisition Research: Theoretical and Methodological Issues
Susan Gass and Jacquelyn Schachter, Editors

Monographs on Research Methodology

Yule • Referential Communication Tasks
Markee • Conversation Analysis

Schachter/Gass • Second Language Classroom Research: Issues and Opportunities

Birdsong • Second Language Acquisition and the Critical Period Hypothesis

In this chapter, we consider the relation between the age at which the naturalistic acquisition of a second language (L2) begins, and the accuracy with which the L2 is pronounced. Quite clearly, earlier is better as far as L2 pronunciation is concerned. However, the widely accepted critical period hypothesis does not appear to provide the best explanation for this phenomenon.

**INTRODUCTION**

Although it is widely agreed that "earlier is better" as far as the pronunciation of an L2 is concerned, there is disagreement as to the exact nature of the relation between the age of L2 learning and degree of foreign accent, as well as the cause(s) of foreign accent (see Singleton, 1989, for a review). Long (1990) concluded from a review of previously published studies that an L2 is usually spoken without accent if learning begins by the age of 6, with a foreign accent if learning begins after the age of 12, and with variable success between the ages of 6 and 12. Patkowski (1990) concluded that the dramatic difference he noted in the foreign accents of participants who had first arrived in the United States before versus after the age of 15 was due to the passing of a critical period, which he defined as an "age-based constraint on the acquisition of full native fluency" in an L2. Indeed, Patkowski claimed that individuals who begin learning an L2 before versus after the critical period differ in a "fundamental, qualitative way" (p. 74).
The critical period hypothesis (CPH) is widely viewed as providing an explanation for why many individuals speak their L2 with a foreign accent. The end of a critical period for speech is usually associated with some sort of neurological change (e.g., lost plasticity, hemispheric specialization, or neurofunctional reorganization) that is thought to arise as the result of normal maturation (e.g., Lamendella 1977; Lenneberg, 1967; Patkowski, 1990; Penfield & Roberts, 1959; Scovel, 1969, 1988). Such a neurofunctional change(s), which might be expected to occur at roughly the same chronological age in many individuals, could conceivably affect the processing and storage in long-term memory of information pertinent to the L2 (e.g., Genesee et al., 1977; Lenneberg, 1967; Patkowski, 1990; Penfield & Roberts, 1959; Scovel, 1969, 1988). The CPH seems to imply that some aspect(s) of the capacity that permits children to learn to pronounce their L1 accurately is reduced or lost beyond the critical period.

Patakowski's (1990) conclusion that a critical period exists for speech learning was based on a pattern of empirical data that has not been replicated in two recent studies. Flege, Munro, and MacKay (1995) examined the production of English sentences by 240 native speakers of Italian who immigrated to Ottawa, Canada, between the ages of 2 and 23. Given Patakowski's (1990) admonition that the CPH can be properly evaluated only by considering participants who have reached their ultimate attainment in L2 pronunciation under optimal learning conditions, Flege et al. (1995) recruited participants who had been living in Canada for at least 15 years at the time they were tested. In fact, the native Italian participants had lived in Ottawa for 32 years on average; most of them indicated that they spoke English more than Italian.

The 240 native Italian participants' productions of five short English sentences (e.g., The red book was good), along with those of a control group of 24 native English participants, were digitized and then presented randomly to native speakers of English from Ontario. These listeners rated the sentences they heard for overall degree of perceived foreign accent using a continuous scale. Figure 5.1 shows the mean ratings obtained for the 264 participants. As expected, the native English participants received higher ratings than most native Italian participants, whose ratings decreased systematically as age of arrival (AOA) increased. Importantly, there was no discontinuity in the ratings at an AOA of 15 years, or at any other AOA. The straight line fit to the data obtained for the 240 native Italian participants accounted for 71.4% of the variance in the ratings accorded their sentences ($p < .01$). (Language use factors accounted for roughly 15% of additional variance; see Flege et al., 1995). A subsequent study by Yeni-Komshian, Flege, and Liu (1997) that employed a similar design also yielded a near-linear relation between AOA and degree of foreign accent in a population of 240 Korean–English bilinguals living in the United States.

We can only speculate on the cause of the important difference in the results obtained by Flege et al. (1995) and Yeni-Komshian et al. (1997), on the one hand, and by Patakowski (1990), on the other hand. The difference was probably not due to differences in average length of residence in an English-speaking country of the participants who were studied—roughly 32 years for the Flege et al. (1995) participants, 20 years for the Patakowski (1990) participants, and 15 years for the Yeni-Komshian et al. (1997) participants. It is conceivable, however, that one or more of the following other factors contributed to the observed difference: heterogeneity of the nonnative groups that were studied (many different native languages (L1s) in the Patakowski, 1990, study, just one L1 in the other two studies); the size of the nonnative population (67 vs. 240); the scaling techniques employed (a 5-point scale by Patakowski, 1990, vs. continuous and 9-point scales in the other two studies); and judges who evaluated the speech materials (trained
English as a Second Language teachers in the Patkowski study vs. untrained).

I think it more likely, however, that the difference was an indirect consequence of the kind of speech materials that were examined. The participants examined by Flege et al. (1995) and by Yeni-Komshian et al. (1997) produced a standard set of sentences, whereas Patkowski (1990) examined 30-second excerpts of spontaneous speech samples that his participants had produced in interviews lasting from 15 to 30 minutes. It is therefore possible that Patkowski’s trained judges were influenced by the nonnative participants’ word choices and grammatical accuracy in addition to differences in pronunciation accuracy. If so, then the results obtained by Patkowski may be indicative of a sharp age-related discontinuity in performance in some linguistic domain other than the phonetic-phonological domain.

In my view, the lack of a nonlinearity in the function relating AOA to degree of foreign accent is inconsistent with the view that a critical period exists for speech learning (see also Bialystok & Hakuta, this volume). There was, however, one aspect of the data obtained by Flege et al. (1995) that was consistent with a CPH. None of the native Italian participants who began learning English after the age of 15 obtained a score that fell within two standard deviations (SDs) of the mean value obtained for the 24 native English control participants, and thus might be deemed to have learned to speak English without a detectable foreign accent. However, the data presented by Bongaerts, Planken, and Schils (1995; see also Bongaerts, Chap. 6, this volume) suggest that certain highly motivated individuals who begin learning their L2 beyond the age traditionally thought to mark the end of a critical period do manage to speak their L2 without foreign accent.

AN INTERACTIONIST PERSPECTIVE

The data just presented pose a problem for the CPH in that they did not reveal a sharp decline in pronunciation accuracy as a function of age. A more general problem with the CPH is that it does not specify the actual mechanism(s) that supposedly deteriorate, or are lost altogether, as the result of maturation. Several possibilities come to mind. For example, neurofunctional change(s) might reduce a person’s ability to add or modify the sensorimotor programs used for producing the vowels and consonants of an L2 (McLaughlin, 1977). Or, change(s) might reduce the ability to establish perceptual representations for new vowels and consonants (Flege, 1995; Rochet, 1995).

Still another problem is that the CPH is not directly testable. This is because factors that might conceivably influence speech learning are inevitably confounded with chronological age, which is the usual surrogate for the state of neurofunctional maturation that is thought to precipitate a lost or slowed ability to learn speech (see Flege, 1987, for discussion). For example, participants’ age of first exposure to an L2 in a predominantly L2-speaking environment may be related to their strength of emotional attachment to the L1-speaking community, their willingness to sound just like members of the L2-speaking culture, or both. To take another example, either length of residence in an L2-speaking environment or chronological age must be confounded in a research design meant to compare groups of participants differing in their AOA in an L2-speaking environment.

As I see it, the most serious problem is that, because of its widespread appeal, the CPH dampens researchers’ enthusiasm for seeking and testing other potential explanations for the ubiquitous presence of foreign accents (as well as age-related declines in other aspects of L2 performance). Other general hypotheses can indeed be formulated. For example, according to what might be called the exercise hypothesis, one’s ability to learn to produce and perceive speech remains intact across the life span, but only if one continues to learn speech uninterruptedly (see Bever, 1981; Hurford, 1991). On this view, foreign accents increase as a function of AOA because as AOA increases, fewer individuals can be found who have never stopped learning speech. However, although it is interesting, the exercise hypothesis may be difficult or impossible to test. It may not be possible to recruit matched groups of participants who have begun to learn some language X, at the same age and under similar circumstances but who differ according to whether other languages were learned between L1 acquisition and the time of first exposure to language X.

According to an unfolding hypothesis, foreign accents are the indirect consequence of previous phonetic development, not the result of lost or attenuated speech learning abilities (Oyama, 1979; see also Elman, 1993, and Marchman, 1993, for a connectionist perspective). For example, the phonetic categories established for vowels and consonants in the L1 may become better defined with age (Flege, 1992a, 1992b) and so become ever more likely to “assimilate” phonetically different vowels and consonants in an L2 (Best, 1995). The unfolding hypothesis predicts that the more fully developed the L1 phonetic system is at the time L2 learning begins, the more foreign-accented the pronunciation of the L2 will be. A problem also exists for the unfolding hypothesis, however. The state of development of the L1 phonetic system is apt to covary with maturation and development (and, of course, chronological age). This means that differentiating the unfolding hypothesis from the CPH may be impossible.

Still another general hypothesis might be called the interaction hypothesis. Weinreich (1953) was apparently the first to suggest that a mutual influence of a bilingual’s two languages on one another is inevitable. If so, it may be impossible for a bilingual to control two
languages in exactly the same way as two monolinguals. Indeed, a number of investigators have suggested that it is not appropriate to assess bilinguals in the same way that one assesses monolinguals (Grosjean, 1982). For example, Cook (1995) observed that divergences from monolingual-defined norms for the L1 or the L2 should not be viewed as a failure, as suggested by Selinker (1972), but as the necessary consequence of "multicompetences" in two languages. Cook (1995) suggested that, in the aggregate, the multicompetences of a bilingual normally exceed the competence of any one monolingual. Mack (1986) noted that although early bilinguals may be quite fluent in both of their languages, the way they process language may differ from that of monolinguals because of a "pattern of linguistic organization that is unlike that of a monolingual" (p. 464; see also Neville, Mills, Lawton, 1992; Weber-Fox & Neville, 1992).

According to the interaction hypothesis, bilinguals are unable to fully separate the L1 and L2 phonetic systems, which necessarily interact with one another. The L1 and L2 systems may, of course, form constrained subsystems that can be activated and deactivated to varying degrees (Paradis, 1993). This is what permits different modes of pronunciation in the L1 and L2. However, according to the interaction hypothesis, the phonetic elements of the L1 subsystem necessarily influence phonetic elements in the L2 system, and vice versa. The nature, strength, and directionality of the influence may vary as a function of factors such as the number and nature of categories established for phonetic elements of the L1 and L2, the amount and circumstances of L1 and L2 use, language dominance, and so on (see e.g., Anisfeld, Anisfeld, & Segui 1969; Cutler, Mehler, Norris, & Segui, 1989; Flege, 1995; Ho, 1986; Macnamara, 1973). The interaction hypothesis leads to a prediction that is not generated by a CPH or any other hypothesis. It predicts that the loss of the L1, or its attenuation through disuse (Grosjean, 1982; Romaine, 1995), may reduce the degree of perceived foreign accent in an L2. In other words, the "less" L1 there is, the smaller will be its influence on the pronunciation of an L2 (Dunkel, 1948).

The interaction hypothesis was tested by Flege, Frieda, and Nozawa (1997) in a study that examined foreign accent in English sentences spoken by native speakers of English and two groups of native Italian participants. The participants in both native Italian groups had arrived in Canada from Italy at an average age of 5 but differed in self-reported use of Italian, 3% on average for the "LoUse" participants versus 33% for the "HiUse" participants. The sentences spoken by the native Italian participants and those spoken by the native English controls were randomly presented to native English-speaking listeners who labeled each sentence as "definitely English" (i.e., definitely spoken by a native speaker of English), "probably English," "probably Italian" (i.e., probably spoken by a native speaker of Italian), or "definitely Italian." The results of two analyses1 yielded two findings that run counter to the CPH. First, sentences spoken by both the HiUse and LoUse participants were found to be foreign accented even though the participants in these groups had learned English as young children and had spoken English for more than 30 years, on the average. The CPH would lead one to expect that childhood learners of an L2 could evade being detected as foreign accented. Second, the HiUse participants were found to speak English with significantly stronger foreign accents than did the LoUse participants. Given that the CPH attributes foreign accent to the state of neurological maturation present at the time L2 learning begins, it would not lead one to expect a difference in L2 performance as a function of amount of L1 use.

The mutual influence of L1 and L2 on one another was also demonstrated in a study by Yeni-Komshian, et al., (1997). The participants for this study were 240 Korean–English bilinguals who had arrived in the United States between the ages of 2 and 23 and had lived in the United States for 15 years on the average (range: 8–30 years). Native English-speaking listeners used a 9-point scale to rate English sentences spoken by the bilingual participants and a control group of 24 English monolinguals for overall degree of foreign accent. In a parallel experiment, native Korean-speaking listeners used a comparable scale to rate Korean sentences spoken by the bilinguals and a control group of 24 Korean monolinguals. When plotted as a function of the bilinguals' AOA in the United States, the functions established for Korean foreign accent in English sentences and for English foreign accent in Korean sentences formed an "X" pattern. The later the Koreans had arrived in the United States, the less accurately they were judged to have pronounced the English sentences and the more accurately they were judged to have pronounced the Korean sentences.

The Yeni-Komshian et al. (1997) study provided evidence that few if any bilinguals pronounce both of their languages without a detectable foreign accent. Just 16 (7%) of the bilinguals received a rating for their production of English sentences that fell within +/- 2 SDs of the mean rating obtained for the 24 English monolinguals. Those who met the criterion had A0As ranging from 1.5 to 8.5 years. A much larger number of the bilinguals, 111 (46%), received a rating for their

1In one analysis, the native English-speaking listeners' judgments were converted to a 4-point rating scale. In another analysis, the correct identifications of native Italian participants' sentences as Italian were counted as hits and incorrect identifications of native English participants' sentences as Italian were counted as false alarms. A-Prime (A') scores representing an unbiased measure of the listener's sensitivity to foreign accent in the native Italian participants' sentences were then calculated from the proportion of hits and false alarms.
proportion of Korean sentences that fell within +/− 2 SDs of the mean obtained for the 24 Korean monolinguals. The bilinguals who met this criterion had AOAs ranging from 8.5 to 22.5 years. However, just one bilingual out of 240, a woman with an AOA of 8.5 years, met the criterion in both English and Korean.

In summary, there is evidence that the two languages spoken by a bilingual interact with one another. If it is true that one continues to learn and refine the phonetic–phonological system of the L1 through childhood and adolescence, then the interaction hypothesis might provide an account for age effects on L2 performance that differs from the one offered by the CPH. Moreover, the interaction hypothesis seems to be more consistent with the observed linear relation between AOA and degree of foreign accent than does the CPH, which leads one to expect a discontinuity.

PRODUCTION AND PERCEPTION

So far we have considered only the overall pronunciation of the L2, that is, degree of foreign accent in sentences. Of course, individuals who learn an L2 often produce particular L2 vowels and consonants inaccurately, which contributes to what is perceived as foreign accent. Some part of nonnatives' divergences from the segmental phonetic norms of the L2 in speech production may arise from an inability to master new forms of articulation. It would be interesting to know, for example, whether native speakers of English who are highly proficient speakers of a Southern Bantu language such as Xhosa are able to master the motorically complex clicks found in that language. If clicks can never be mastered by native English-speaking adults, it would suggest the existence of age constraints on articulatory motor learning.

However, many researchers (e.g., Flege, 1988b) believe that certain speech production errors arise from an incorrect perceptual representation of the properties that specify L2 vowels and consonants. For example, Rochet (1995) used a synthetic /i/-/y/-/u/ continuum of vowels to assess nonnatives' vowel perception. Native Portuguese participants tended to misidentify French /y/ as /i/, whereas native English participants tended to misidentify the same vowel stimuli as /u/. In a repetition task, native Portuguese participants produced /i/-quality vowels when they heard French /y/ tokens, whereas native English participants tended to produce /u/. This finding led Rochet to conclude that some vowel production errors are "the consequence of the target phones having been assigned to an L1 category" (p. 404).

The Speech Learning Model (SLM), e.g., Flege, 1995) posits that the L1 and L2 influence one another, and that this interaction constrains performance accuracy in both languages. According to the SLM, a diversity of factors such as an individual's age of L2 learning and the perceived dissimilarity of L2 sounds from the closest L1 sound(s) determine whether an L2 learner will discern the phonetic differences that may exist between an L2 sound and the closest (nonidentical) sound in the L1. Awareness that a cross-language difference exists, in turn, may precipitate the formation of a new L2 phonetic category. Flege hypothesized that "the production of an (L2) sound eventually corresponds to the properties present in its phonetic category representation" (p. 239). This implies that, for certain L2 learners, the perception of an L2 sound may be more accurate than its production.

The hypothesis that production accuracy is constrained by perceptual accuracy is by no means new. Researchers generally agree that speech perception becomes attuned to the contrastive sound units of a particular language very early in life and that, in time, children's production of speech corresponds to what they have heard. Kuhl and Meltzoff (1996) posited that skilled articulation arises out of language-specific perception. They observed that, for the mature native speaker, information specifying auditory-articulatory relations is "exquisitely detailed . . . as though adults have an internalized auditory-articulatory 'map' that specifies the relations between mouth movements and sound" (p. 2425). They also observed that the formation of memory representations "derives initially from perception of the ambient input and then acts as guides for motor output" (p. 2425). Pisoni (1995) noted that the phonetic contrasts that are produced by talkers are "precisely the same acoustic differences that are distinctive in perceptual analysis," making the relation that exists between speech production and perception "unique" among category systems. Pisoni noted further that although the relation between production and perception is apt to be "complex," it is a nonarbitrary relation that reflects the properties of a "unitary articulatory event" (pp. 22-23).

L1 acquisition research has yielded results that are consistent with the views expressed by researchers such as Kuhl and Pisoni. For example, Kuijper's (1996) work suggested that children's ability to produce and perceive L1 segmental contrasts develops slowly and in parallel through early childhood. However, it is not universally

However, a close relation between motor control and perception may not be a characteristic that is unique to speech. A close relation between production and perception seems to reflect a general characteristic of brain functioning. Churchland (1986) observed that "evolution [has] solved the problem of sensory processing and motor control simultaneously," so that "theories [must] mimic evolution and aim for simultaneous solutions as well" (p. 473). According to Edelman's theory of neuronal group selection (e.g., Edelman, 1989), a "dynamic loop . . . continually matches gestures and posture to several kinds of sensory signals." In Edelman's view, perception "depends upon and leads to action." Motor activity is considered to be "an essential part of perceptual categorization" (54-56).

5. AGE OF LEARNING AND L2 SPEECH
accepted that the same kind of parallelism between production and perception exists in L2 acquisition. In fact, Bever (1981) hypothesized that a critical period for learning speech exists because the development of production and perception will not be closely linked if speech learning occurs after the L1 is firmly established.

Bever (1981) postulated that, during L1 acquisition, a psychogrammar "equilibrates" (or aligns) production and perception. Bever rejected the view that the development of speech perception during L1 acquisition necessarily precedes corresponding developments in speech production, observing that advances in the two domains “leapfrog.” He posited that psychogrammar representations reflect the “conjoint” operation of perception and production, and that it is only through the mediation of such representations that what the child has acquired perceptually can influence production, and vice versa.3

According to Bever (1981), as the L1 phonology is acquired, production and perception are brought into alignment. Use of the psychogrammar will cease once its primary role, which is to align production and perception, has been accomplished. At this point in speech development, Bever hypothesized that speech production and perception become “independent” and the critical period for speech learning ends. Although psychogrammar representations for the L1 might be accessed, the psychogrammar can no longer be used to align production and perception in an L2 learned after the critical period, that is to say, after the L1 phonology has been fully acquired. It is for this reason that L2 learners “often learn to discriminate sounds … they cannot distinctively produce” (Bever, 1981, p. 196).

Bever’s version of the CPH is valuable because of its specificity. However, it is not compatible with the results obtained in recent studies that examined L2 segmental production and perception. These studies suggest that, as in L1 acquisition, the production and perception of L2 vowels and consonants may “align” with one another. However, before turning to these segmental studies, I first review a number of related studies dealing with sentence production and perception.

Sentence-Level Studies

By extension, Bever’s (1981) psychogrammar hypothesis might lead one to expect that a nonnative’s ability to produce and comprehend sentences in their L2 will be unrelated. However, two studies suggest otherwise.

5. AGE OF LEARNING AND L2 SPEECH

Oyama (1973; see also Oyama, 1982a, 1982b) tested 60 Italian men living in New York City who had arrived in the United States between the ages of 6 to 20 and had lived there for 5 to 18 years. She assessed the participants’ degree of perceived foreign accent by having English-speaking listeners rate paragraph-length speech samples. Sentence comprehension was assessed by having the participants repeat as many words as possible in a set of English sentences presented in noise. The foreign accent ratings and the scores from the comprehension test (i.e., the total number of words that could be repeated) for the individual participants were not available for reanalysis. However, when the mean values obtained for six subgroups of the participants (defined on the basis of length of residence and AOA in the United States) was examined, a significant correlation ($r = 0.818$) was obtained. This indicated that the better the native Italian participants pronounced English, the better they were at comprehending English sentences in noise.

The sentences examined by Oyama (1973) included some words that were predictable from context (e.g., *Shepherds seldom lose their sheep*). The scores Oyama obtained were therefore likely to have been influenced to some extent by the participants’ higher order knowledge of English. Meador, Flege, and MacKay (1997) recently replicated and extended the Oyama study. To obtain scores that more closely reflected the bottom-up processing of vowels and consonants, semantically unpredictable English sentences (e.g., *The blond dentist ate the heavy bread*) having a single syntactic form (NP-V-NP) were examined. Figure 5.2 shows the results obtained for 54 native Italian participants with a mean age of 48. These participants had arrived in Canada between the ages of 3 and 23 and had lived there for an average of 34 years. The more accurately the participants pronounced English sentences (as rated by native English-speaking listeners), the larger the number of words they were able to repeat, $r = 0.646, df = 52, P < 0.001$.

The correlation just reported ($r = 0.646$) may actually have underestimated the relation between the participants’ ability to produce and perceive the vowels and consonants in the sentences. Repeating the words of a sentence presented in noise requires that words, or parts of words, be held in memory while additional information is processed. Individual differences in phonological short-term memory (PSTM) might therefore be expected to influence performance on a sentences-in-noise task. The participants’ PSTM was assessed by having them repeat nonwords formed by concatenating two to five Italian CV syllables. When the variation in the PSTM scores was partialled out, the correlation between the participants’ degree of

---

3It is presumably the existence of psychogrammar representations that permits a child to know that /ft/ and /ft/ are two ways to say the same thing (one her own way and the other an adult’s way).
of Words Repeated

5. AGE OF LEARNING AND L2 SPEECH

As mentioned earlier, Bever (1981) hypothesized that speech production and perception develop independently during L2 acquisition because the psychogram is no longer used to align production and perception after the end of a critical period. If this hypothesis is correct, then one would not expect to observe correlations between measures of post-critical period L2 learners' production and perception of L2 vowels and consonants. However, the results obtained in recent studies do show significant, albeit modest, correlations.

Segmental Studies

As mentioned earlier, Bever (1981) hypothesized that speech production and perception develop independently during L2 acquisition because the psychogram is no longer used to align production and perception after the end of a critical period. If this hypothesis is correct, then one would not expect to observe correlations between measures of post-critical period L2 learners' production and perception of L2 vowels and consonants. However, the results obtained in recent studies do show significant, albeit modest, correlations.

Vowels

A study by Flege, Bohn, and Jang (1997) provided evidence that a relation exists between late bilinguals' production and perception of L2 vowels. The participants were 20 native speakers each of English, German, Spanish, Korean, and Mandarin. The 80 nonnative participants were first exposed to English on a regular basis when they arrived in the United States as adults; they had lived in the United States for an average of 4.0 years (range: 0.2–23 years) at the time of testing. The participants read a list of consonant-vowel-consonant (CVC) English words containing the vowels /i:/, /ɪ:/, /ɛ/ and /æ/. Later, they identified the members of two continua containing the same vowels. The perceptual stimuli used in two-alternative forced-choice identification experiments consisted of synthetic vowels that ranged from /i:/ to /ɪ:/ in one continuum, and from /ɛ/ to /æ/ (between /ɛ/ and /æ/) in the other continuum. In both continua, spectral quality (F1 and F2 frequency) was varied in 11 steps and vowel duration was varied orthogonally in three steps.

One method used to assess vowel production accuracy was to measure the size of the spectral (F1, F2) differences that the participants produced between /i:/–/ɪ:/ and /ɛ/–/æ/. The native English participants...
relied primarily on spectral (F1, F2) variation to identify vowels as /i/ versus /ɪ/ (or as /ɛ/ versus /æ/). Many nonnative participants, on the other hand, relied mostly or even entirely on vowel duration, perhaps because they did not have two separate, spectrally defined representations for the perceptual continuum endpoints. Thus, one way used to assess the participants' accuracy in perceiving English vowels was to determine the extent to which their identification responses changed from one response category to the other as a function of the 11-step spectral manipulation in each continuum.

Figure 5.3a shows the relation between the 80 nonnative participants' vowel production and perception accuracy for English /i/-/ɪ/. The percentage change in /i/ responses that occurred as a result of the spectral manipulation in the synthetic vowel stimuli is shown on the x-axis. The y-axis shows the magnitude of the spectral differences between /i/-/ɪ/ that the participants produced (i.e., the Euclidean distance between the values measured in two vowels when plotted in a 2-dimensional bark-difference space, B2-B2 vs. B1-B0.) The more the participants changed their identifications as formant frequencies were changed, the larger were the spectral differences between /i/-/ɪ/ that they produced, $r = 0.529$, $p < 0.01$. Similarly, as shown in Fig. 5.3b for /ɛ/-/æ/, the greater the increase in /æ/ responses in the perception experiment, the larger the spectral difference that the participants produced between /ɛ/-/æ/, $r = 0.523$, $p < 0.01$.

Both production--perception correlations just reported were significant, but they were modest in size. This does not undermine the view that production accuracy is constrained by perception accuracy in L2 acquisition. Indeed, it is just what one expects if accuracy in perception is a prerequisite for accuracy in production but does not guarantee it. Inspection of the individual participant data in Fig. 5.3 reveals that the participants who showed a large (and thus English-like) shift in judgments as spectral quality varied showed a wide range in production accuracy. On the other hand, most of the participants who showed little perceptual effect of spectral quality also produced little spectral difference between the English vowels.

There were, of course, individual exceptions to the general pattern. Thus, the following analysis was carried out to determine if the participants with inaccurate perception also tended to produce vowels inaccurately. The nonnative participants were assigned to one of three subgroups based on their performance in the identification experiment. Those who showed more than an 80% decrease in /i/ as F1 values increased in the /i/-/ɪ/ stimuli ($n = 18$) were designated the "accurate" perceivers. Those who showed shifts of 10% to 79% ($n = 22$) were designated the "moderately accurate" perceivers; and those who showed shifts of less than 10% in the expected direction ($n = 40$) were
Flege (1993) examined Chinese participants' production and perception of /t/ and /d/ in the final position of English words. This phonetic contrast was of interest because Chinese words are not differentiated by word-final obstruents that differ in voicing. The study focused on vowel duration. Native speakers of English make vowels longer before /d/ than /t/ in words such as bead versus beat. If asked to identify an ambiguous word-final stop as /d/ or /t/, a relatively long vowel gives rise to the perception of /d/ by native speakers of English. Thirty of the Chinese participants who participated were "late" bilinguals who were first exposed to English on a regular basis as adults when they arrived in the United States; nine others were "early" bilinguals who had arrived in the United States prior to the age of 10. Both the late and early bilinguals made vowels significantly longer in English words ending in /d/ than /t/ (p < .01), but the size of the late bilinguals' contrasts were significantly smaller than those produced by the native English controls and the early Chinese-English bilinguals (p < .01).

A parallel perception experiment by Flege (1993) assessed the Chinese-English bilinguals' use of vowel duration as a cue to the voicing feature in stops. Both 17-member perceptual continua that were developed consisted of naturally produced English CVC words in which the original vowel durations were altered in such a way that native English controls heard a word ending in /d/ (for stimuli with the longest vowels) or words ending in /t/ (for stimuli with the shortest vowels). The method of adjustment was used. In one session making use of a beat-bead continuum, the participants were asked to choose the member of the continuum that represented the best example of beat. In a second session using the same continuum, they were asked to choose the best instance of bead. In two sessions using a bat-bad continuum, the participants were asked to choose the best examples of bat or bad.

Both groups of Chinese-English bilinguals examined by Flege (1993) chose stimuli with significantly longer vowels as the best instances of /d/-final compared to /t/-final English words. However, the magnitude of the perceptual effect was significantly smaller for the late bilinguals than for the native English controls or for the early bilinguals (p < .01). The bilinguals' perceptual and productive use of vowel duration thus seemed to parallel one another. Indeed, a significant correlation was obtained between the size of vowel duration differences that the bilinguals produced in /t/-final versus /d/-final words and the size of the vowel duration differences observed between the stimuli preferred as the best instances of /t/-final versus /d/-final words, r = .535, p < .01.

Recent studies examining the voice onset time (VOT) dimension in the production and perception of word-initial English stop consonants also suggest that production and perception are related at a segmental level during the acquisition of L2 speech. This work (Flege and Schmidt, 1995; Schmidt and Flege, 1995) examined 40 native speakers of Spanish who came to the United States as young adults. The participants judged the members of two synthetic continua. The stimuli in the continua were heard by native English control participants as /bi/, /pi/, or exaggerated /pi/ (i.e., a stop with too much aspiration) depending on VOT. One continuum consisted of short-duration CVs that simulated a fast speaking rate; the other consisted of long-duration CVs that simulated a slower rate of speech.

The participants rated randomly presented members of both continua for goodness as instances of the English /p/ category. As shown in Fig. 5.4, native English control participants gave low ratings to the stimuli that had VOT values that were shorter than is typical for English /p/. As VOT increased, so too did their goodness ratings; but as VOT values increased beyond values typical for English, the goodness ratings began to decrease systematically. The native English participants exhibited "internal category structure" in that, for them, some stimuli were better examples of /p/ than others.

The VOT value of the stimulus that received the highest rating, called the preferred VOT value, was determined for each participant. A significant correlation was found to exist between the VOT values that the native English participants produced and their perceptually preferred VOT values, r = .536, p < .01 (see Newman, 1996, for similar results). That is, the native English participants who produced /p/ with relatively long VOT values tended to prefer stimuli having relatively long VOT values. This evidence of alignment does not contradict Bever's (1981) hypothesis, for the alignment could have been established during L1 acquisition, prior to the ending of a critical.
The results obtained for native Spanish participants in other research do run counter to Bever's hypothesis, however. In Spanish, /p/ is produced with short-lag VOT values rather than with the long-lag VOT values typical for English. Native speakers of Spanish who learn English in adulthood have been observed to produce voiceless English stops such as /p/ with VOT values ranging from Spanish-like short-lag VOT values to values that match or even exceed the long-lag VOT norm for English.

Flege and Schmidt (1995) determined the overall degree of foreign accent in English sentences spoken by 40 Spanish late bilinguals, assigning the 20 participants with the lowest ratings to a nonproficient group and the 20 with the highest ratings to a relatively proficient subgroup. The correlation between the VOT values produced by the proficient participants and their perceptually preferred VOT values was significant, \( r = .489, p < .01 \), whereas the correlation observed for the nonproficient native Spanish participants was nonsignificant, \( r = -.004, p > .10 \). This finding suggests that as nonnative adults gain proficiency in an L2, their production and perception align.

To summarize the evidence presented so far, it appears that adult learners' overall pronunciation of their L2 and their ability to comprehend it are related. However, the ability to perceptually gauge degree of accent in L2 sentences may develop more rapidly, or to a greater extent, than the ability to pronounce L2 sentences.

At the segmental level, modest correlations have been found to exist between production and perception accuracy. Segmental production and perception do not appear to develop independently as hypothesized by Bever (1981). One possible explanation for why the segmental production-perception correlations observed so far have been modest is that not all participants who adapt their perception to conform to the sound pattern of the target L2 make a comparable adaptation in production. If perception "leads" production in L2 acquisition, then the modest correlations that have been observed are just what one would expect. To use a term coined by Bever (1981), certain adult learners may not yet have transported what they learned about the perception to the domain of speech articulation.

Other explanations might also be advanced for the modest size of the correlations observed. For example, although segmental phonetic contrasts are based on multiple dimensions, most published studies have focused on a single dimension. It is possible that some dimension(s) other than the one examined in a study have undergone change as the result of learning in one or both domains. Or, a stronger underlying relation between production and perception may have been obscured by measurement error or some inadequacy in experimental design. For example, the speech production samples may have represented fast-rate speech, whereas the speech perception data may have represented careful speech produced at a slower rate.

Still another possible explanation for why the observed L2 production-perception correlations tend to be significant but weak is that the most meaningful perceptual variable has not yet been examined. According to the SLM (e.g., Flege, 1995), category formation exerts a powerful influence on L2 learners' accuracy in producing L2 vowels and consonants. The notion of category formation implies a discontinuity in performance. This being the case, it might prove more fruitful to compare the segmental level performance of participants who have versus have not formed a category than to compare, for example, groups of participants differing in overall L2 proficiency or AOA in an L2-speaking environment.

There is some preliminary evidence to support the hypothesis that production accuracy is related to category formation. Flege, MacKay, and Meador (1998) elicited the production of 11 English vowels in two ways. One of the two elicitation methods was thought likely to require
the presence of English vowel categories. The participants examined were three groups of 18 native Italian speakers each who had arrived in Canada at average ages of 7, 13, and 19 respectively, plus a group of English controls. These participants were all long-time residents of Ottawa, Canada, with a mean age of 48. The native Italian participants had lived in Canada for an average of 34 years at the time of testing, and estimated speaking Italian 31% of the time, on average.

The participants in the Flege et al. (1998) study were given a list of the CVC words so that they could read as well as hear the vowels that were said. For each vowel of interest, the participants first repeated a sequence of four real words containing a single vowel (example: read, deed, head, head) after hearing the words via a loudspeaker. After hearing the same four-word sequence a second time, the participants inserted the vowel found in all four words (/i/, in the example given) into a /b__d/ frame. After a third and final presentation of the four-word sequence, they inserted the /bVd/ nonword (where V = the vowel common to all four real words) into a carrier phrase (I say again and again). After digitization, productions of each vowel were randomly presented in separate blocks (one for each vowel of interest) to native speakers of English from Ottawa for goodness ratings. The identity of the intended vowel in each block was always known beforehand to the listeners.

Figure 5.5a shows the goodness ratings obtained for four English vowels that have a phonetically different counterpart in Italian. Separate mean ratings are shown for the vowels that were produced in real words (i.e., the vowels in the last words of the four-word sequences) and the vowels spoken in a /b__d/ frame (i.e., the nonwords inserted into the carrier phrase). The native Italian participants' accuracy in producing /i e o u/ increased somewhat as a function of AOA. However, there was little difference in the accuracy with which the four groups of participants produced these vowels in words versus nonwords, so the Group x Vowel interaction in the ANOVA examining the goodness ratings for /i e o u/ was nonsignificant, $F(3,68) = 2.4, p > .05$.

As shown in Fig. 5.5b, however, a different pattern of results was obtained for four English vowels that are unlike any vowel in the Italian inventory, namely /a e u a/. The native Italian participants with AOA of 7 and 13 did not differ significantly from the native English controls when producing these vowels in real words ($p > .10$).

5 Another group of native Italian participants with an AOA of 7 years was made up of individuals who seldom spoke Italian. As might be expected from the interaction hypothesis, they performed more like the native English controls than did participants matched for AOA who reported speaking Italian more often.

6 The vowels in the real word condition shown in Fig. 5.5 were those in beard, bid, bed, bad, bade, hood, hooded, bird.

FIG 5.5a. Mean goodness ratings obtained for English vowels with a counterpart in Italian.

However, unlike the participants in the other two groups, they produced /æ u æ/ significantly less accurately in nonwords than in real words, and significantly less accurately than the native English controls ($p < .01$). This led to a significant two-way interaction, $F(3,68) = 15.8, p < .001$.

To produce the /æ u æ/ accurately in nonwords, the native Italian participants had to identify the vowel heard in four real words, hold that vowel in working memory for an interval of time, and then produce the represented vowel in /b__d/ context, thereby forming a nonword. It is unlikely that the difficulty experienced by the native Italian

7 The lack of an effect of Condition for the native Italian participants with an average AOA of 19 years can be attributed to the fact that they produced /æ u æ/ so inaccurately in real words that a further reduction in accuracy in the nonword condition was not realistically possible.
participants with AOA of 7 and 13 years in producing /i e u o/ was due to memory limitations. First, they did not produce English vowels that have a counterpart in Italian (namely, /i e o u/) less accurately in nonwords than in words. Second, the participants’ phonological STM was evaluated by having them repeat nonwords formed by concatenating two to five Italian CV syllables. The participants with AOA of 7 and 13 years did not differ in their nonword repetition ability from any other group. The fact that these participants produced /i e u o/ as accurately as the native English participants in the real-word condition suggests that they were motorically able to articulate these vowels accurately. It is likely, therefore, that the native Italian participants’ inaccurate production of /i e u o/ in the nonwords was due either to a lack of long-term memory representations for these vowels, or to representations that did not conform as closely to the /i e u o/ tokens as was the case for the native English controls.

Two other recent studies related to the issue of category formation examined the production and perception of English /p/ by 10 Spanish late bilinguals (Flege, Schmidt & Wharton, 1996; Schmidt & Flege, 1996). The participants rated the members of slow-rate and fast-rate VOT continua (see earlier sections) for goodness as instances of English /p/. As was shown previously in Fig. 5.4, the rating function obtained from native English controls differed as a function of the simulated speaking rate in the two sets of VOT stimuli. More specifically, the English controls gave different goodness ratings to stimuli with VOT values of 50 to 125 msec for English and Spanish) for these vowels, or to representations that did not conform as closely to the /i e u o/ tokens as was the case for the native English controls.

Given that the short-lag /p/ of Spanish shows little variation in VOT as a function of speaking rate (Schmidt & Flege, 1995), a question of interest was whether the Spanish late bilinguals would also show evidence of rate-dependent processing. They might show evidence of internal category structure (the rise–fall of ratings seen in Fig. 5.4), simply by recognizing that English /p/ has longer VOT values than Spanish /p/, and that VOT values that extend beyond the norm for English (the “exaggerated” /p/ tokens) do not occur in human languages. However, it seemed unlikely that they would also show rate-dependent processing (i.e., a systematic shift in their goodness rating as a function of speaking rate) if they did not have a long-term memory representation for English /p/.

Figure 5.6a shows the mean goodness ratings that were obtained for the 4 Spanish late bilinguals (out of 10) who produced English /p/ with the Spanish-like short-lag VOT values ranging from 13 to 18 msec. These four participants showed little if any effect of the speaking rate manipulation when rating the stimuli for goodness as instances of English /p/.

As shown in Figure 5.6b, on the other hand, the four participants who produced English /p/ with English-like long-lag VOT values ranging from 41 to 68 msec did show evidence of rate-dependent processing.

One might speculate that they were able to produce English /p/ accurately because they had established a phonetic category for it.
Additional research of a longitudinal nature will be needed to evaluate this interpretation before it can be accepted, of course. It would also be valuable to determine if teaching native Spanish participants to accurately produce long-lag stops in English would precipitate rate-dependent processing (see Bradlow, Pisoni, Yamada, & Tohkura, 1996). Were such a finding obtained, it would undermine the claim that perception accuracy precedes, and limits, L2 segmental production accuracy.

SUMMARY

In the first section, we presented the results of recent studies examining the relation between age of learning an L2 and degree of foreign accent in the L2. The finding that production accuracy declines linearly with age is inconsistent with the view that foreign accents occur as the result of the passing of a maturationally defined critical period. Then we described alternate hypotheses that might be advanced to account for the fact that earlier is better in regard to the pronunciation of an L2. L2 pronunciation accuracy may decline, not because one has lost the ability to learn to pronounce, but because one has learned to pronounce the L1 so well. The results presented suggest that one's accuracy in pronouncing an L2 varies as a function of how well one pronounces the L1, and how often one speaks the L1.

We next considered a CPH presented by Bever (1981), one that was more specific (and thus testable) than most CPHs. Bever proposed that a critical period for speech learning ends when humans lose the capacity for adapting their production of sounds (vowels and
consonants) to conform to their perceptual representations of the sounds. Empirical evidence was presented that disconfirmed this hypothesis. Other studies were cited showing that modest albeit significant correlations exist between the accuracy with which vowels and consonants are produced and perceived.

According to the SLM (Flege, 1995), the likelihood that L2 learners will establish new categories for L2 vowels and consonants decreases as the age of exposure to an L2 being learned naturalistically increases. It is also hypothesized that the likelihood of category formation for a particular L2 vowel or consonant is related directly to its degree of perceived phonetic dissimilarity to the closest L1 vowel or consonant. In the fourth section, we presented results that were consistent with these hypotheses and with the view (Flege, 1995) that L2 segmental production accuracy is limited by the accuracy of the perceptual representations that are developed for L2 vowels and consonants.

As mentioned earlier, one hypothesis that warrants additional testing is that late bilinguals sometimes establish phonetic category representations for English sounds not found in the L1. Those who do establish a phonetic category representation for an L2 sound may be more accurate in producing it than those who do not. If this hypothesis is correct, then future research examining the relation between production and perception should use discrete tests of category formation instead of (or in addition to) continuous tests of the perception of particular perceptual "cues" (e.g., vowel duration, VOT). Such a research strategy may yield greater insight into the perception-production relation than has yet been obtained in segmental studies.

One caveat should be offered before closing. We are not sure at present what precipitates category formation. It would be prudent, therefore, to remain open to the possibility that category formation is precipitated by the discovery of articulatory means for producing a novel L2 phonetic contrast, and that perceptual fine tuning is mediated by an implicit knowledge of how L2 sounds are produced. One serious obstacle we must face in pursuing these and other important questions is that, at present, an accepted method does not exist with which to test for the formation of new phonetic categories (but see Flege, 1998). To implement the research strategy proposed here, a reliable method must be developed.

Another promising avenue for future research is to study the effects of training. Work by Yamada and Bradlow (Bradlow et. al., 1996; Yamada, Tohkura, Bradlow & Pisoni, 1996) showed that training-induced improvements in perceiving a novel L2 phonetic contrast lead to more accurate production of the L2 contrast in the absence of speech production training. It would be valuable to replicate and extend these findings. Also, it would be worthwhile to determine if training nonnatives to produce new or difficult contrasts will lead to a concomitant improvement in their perception of the L2 contrasts in the absence of perceptual training.

ACKNOWLEDGMENTS

This research was supported by grants DC00257 and DC02892 from the National Institute for Deafness and Other Communicative Disorders. I thank the following individuals for their important contribution to work presented here: Elaina Frieda, Hua Liu, Diane Meador, Ian R. A. MacKay, Murray J. Munro, Takeshi Nozawa, and Grace Yeni-Komshian.

REFERENCES


5. AGE OF LEARNING AND L2 SPEECH


