CHAPTER 5

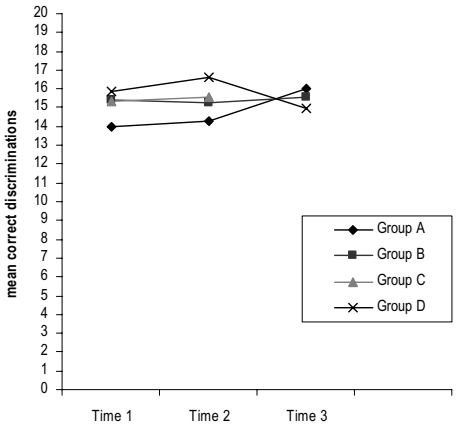
RESULTS

Chapter 5 presents the results obtained for the AX and imitation tasks (Sections 5.1 and 5.2, respectively). It should be recalled that the 281 cross-sectional subjects' answers to the auditory discrimination task were investigated by means of nonparametric tests, according to data screening results. As for the imitation task, based on the preliminary findings stated in 4.4.2, this chapter reports on two studies that were further undertaken in order to examine learners' production data.

5.1. Auditory discrimination task

To test for the effect of the research variables on the learners' answers to the AX task, separate Kruskal-Wallis tests (or Mann-Whitney *U* tests, in the case of two-level independent variables) were performed on the total correct responses (dependent variable) with age of FL learning (AOL), exposure, dominant L1(s), and gender as factors (one factor at a time). Figure 5.1 below shows the mean correct discriminations obtained for the AX task as a function of AOL and exposure to the FL.

Figure 5.1. Overall correct discriminations.

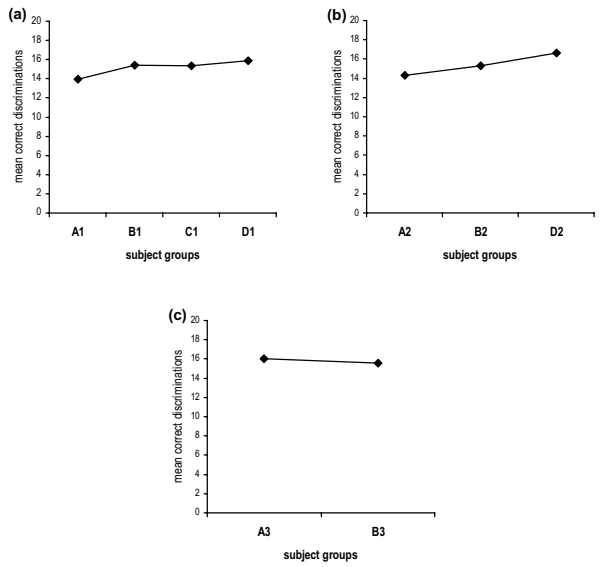


5.1.1. Effect of onset age of FL learning

When age of FL learning was considered, significant differences were found among A1, B1, C1, and D1 (χ^2 16.959, *df* 3, $p < .05$) and among A2, B2, and D2 (χ^2 13.267, *df* 2, $p < .05$). But AOL did not have a significant effect on the discrimination scores obtained by A3 and B3 (*U* 456, *Z* –1.112, $p > .05$).

To explore the significant differences found among groups with various starting ages of FL learning, Mann-Whitney *U* tests were carried out on the total number of correct discrimination scores as the dependent variable. These analyses yielded significant differences between A1 and B1 (*U* 220.5, *Z* –3.015, $p < .05$), A1 and C1 (*U* 183.500, *Z* –2.614, $p < .05$), and A1 and D1 (*U* 335, *Z* –3.934, $p < .05$), all in favour of the older child, adolescent and adult groups. That is, the 8-year-old beginners discriminated significantly more poorly than the other age groups when they had 200 hours of exposure (*M* = 13.97 for A1 vs. 15.39, 15.36, and 15.90 for B1, C1, and D1, respectively). No other two-group comparison reached significance with the same amount of exposure (see Figure 5.2a).

Figure 5.2. Mean correct discriminations. Factor: age of onset of FL learning.



Mann-Whitney *U* tests were also significant for the following pairwise comparisons when experience in the FL amounted to 416 hours: A2 and B2 (*U* 367, *Z* –2.080, $p < .05$), A2 and D2 (*U* 58, *Z* –3.289, $p < .05$), and B2 and D2 (*U* 74, *Z* –2.340, $p < .05$). Between the two adolescent groups, older adolescents (B2) performed significantly better than younger adolescents (A2). And adult learners (D2) discriminated English sounds better than subjects who were 12.99 (A2) and 15.01 (B2) years old on average at Time 2 (see Figure 5.2b).

On the whole, comparisons across subjects who started to learn the FL at different ages (and matched for amount of exposure) showed a significant effect of onset age of FL learning in the initial stages of FL acquisition and after approximately 2–3 years of added exposure on "late" starters. Therefore, older learners obtained significantly higher discrimination scores than younger child beginners. However, age differences among groups became nonsignificant at the final stage of data collection, when 8-year-old starters obtained slightly higher correct discrimination scores than 11-year-old beginners (M = 16.04 and 15.58 for A3 and B3, respectively) (see Figure 5.2c). Table 5.1 below summarises all the group comparisons conducted on the overall discrimination scores and with onset age of FL learning as a factor.

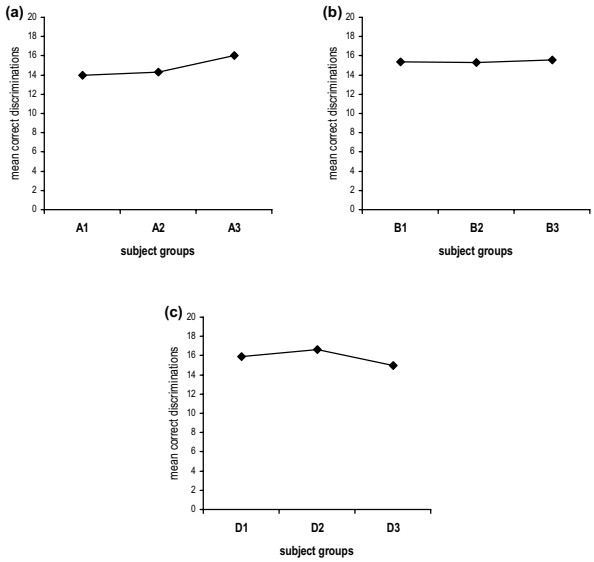
Table 5.1. Summary of comparisons carried out with AOL as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). Each group's mean correct discrimination scores appear in parentheses.

Factor: ***onset age of FL learning***
D.V.: overall correct discrimination scores in AX task
A1 (13.97) – B1 (15.39) – C1 (15.36) – D1 (15.90)*
A1 < B1*
A1 < C1*
A1 < D1*
B1 – C1 n.s.
B1 – D1 n.s.
C1 – D1 n.s.
A2 (14.28) – B2 (15.28) – D2 (16.60)*
A2 < B2*
A2 < D2*
B2 < D2*
A3 (16.04) – B3 (15.58) n.s.

5.1.2. Effect of exposure

Kruskal-Wallis analyses with amount of exposure as a factor were significant in groups of subjects who started to learn the FL at 8 years of age (groups A1, A2, and A3) (χ^2 19.823, *df* 2, $p < .05$). Mann-Whitney *U* tests revealed that A3, the 8-year-old starters with the highest amount of exposure in the research design (i.e. 726 hours), discriminated English sound contrasts and distractors significantly better than the other two groups matched for AOL, but with less instruction: A1 with 200 hours (*U* 144.5, *Z* –4.121, $p < .05$) and A2 with 416 hours (*U* 223.5, *Z* –3.698, $p < .05$). The differences in mean scores between A1 and A2 (*M* = 13.97 and 14.28) were nonsignificant (*U* 499, *Z* –.309, s .758, $p > .05$) (Figure 5.3a).

Figure 5.3. Mean correct discriminations. Factor: exposure to the FL. Results for D3 are indicative only.



Learners who were first exposed to the target language at the age of 11 hardly varied in their total correct discrimination scores, as they received more formal instruction in the FL (*M* = 15.39, 15.28, and 15.58, for B1, B2, and B3, respectively).

Kruskal-Wallis analyses confirmed that there were no significant differences among groups in the discrimination of pairs in the AX task (χ^2 1.090, *df* 2, s .580, $p > .05$) (Figure 5.3b).

For 14-year-old starters, though not statistically analysed, the correct scores that both C1 and C2 obtained showed that they discriminated English sounds at very similar rates with 200 hours and 416 hours of exposure ($M = 15.36$ and 15.57).

Adult learners obtained higher correct discrimination scores halfway through their learning – i.e. 416 hours ($M = 16.60$ for D2 vs. 15.90 for D1 and 15.00 for D3[69]). However, these scores were not significantly higher than those of the other group with 200 hours of instruction in English according to a Mann-Whitney *U* test (U 184.5, Z –1.234, $p > .05$) (Figure 5.3c).

Therefore, an increase in the amount of exposure resulted in significantly higher discrimination scores only for 8-year-old beginners, as displayed in Table 5.2 that summarises the group comparisons made with exposure as a factor.

Table 5.2. Summary of comparisons carried out with exposure as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). Each group's mean correct discrimination scores appear in parentheses.

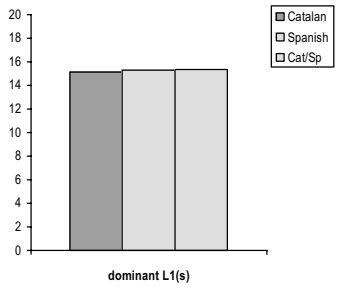
Factor: ***exposure to English***   *D.V.:* overall correct discrimination scores in AX task
A1 (13.97) – A2 (14.28) – A3 (16.04)*   A1 – A2 n.s.   A1 < A3*   A2 < A3*   B1 (15.39) – B2 (15.28) – B3 (15.58) n.s.   D1 (15.90) – D2 (16.60) n.s.

[69] Recall that due to the very few Ss comprising D3, results for that learner group were not included in the statistical analyses conducted. Instead, results for D3 have been considered as an indication/example of the performance on the AX task by a (hypothetical) larger sample of Ss.

5.1.3. Effect of dominant L1(s)

Averaged over the 281 Ss, mean correct discrimination scores were 15.12 for Catalan dominant speakers, 15.32 for Spanish dominant speakers, and 15.36 for Catalan/Spanish balanced bilinguals (see Figure 5.4). The similarity in scores was further demonstrated by means of a Kruskal-Wallis analysis, which showed that the scores were not significantly different (χ^2 1.211, *df* 2, $p > .05$). In addition, there was no consistent pattern of a specific dominant L1 speaker group obtaining higher (or lower) discrimination scores than the other two L1 subgroups across the same AOL and/or amount of exposure.

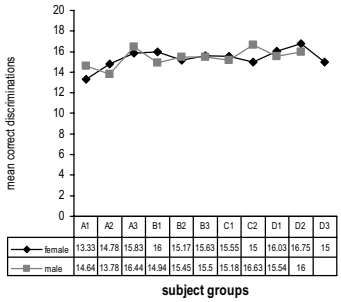
Figure 5.4. Mean correct discrimination scores for Catalan dominant speakers, Spanish dominant speakers, and Catalan/Spanish balanced bilinguals (averaged over the 281 Ss).



5.1.4. Effect of gender

Like dominant L1(s), the variable of gender did not have a significant effect on the correct discrimination scores obtained by male and female subjects (M = 15.12 and 15.43 averaged over male speakers and female speakers, respectively) (U 8645, Z –1.289, $p > .05$). Likewise, no consistent pattern was found in a specific age group or across groups with the same amount of exposure (see Figure 5.5).

Figure 5.5. Mean correct discriminations for male and female subjects in each learner group. Results for C2 and D3 are indicative only.



5.1.5. Effects of research variables on specific sound contrasts

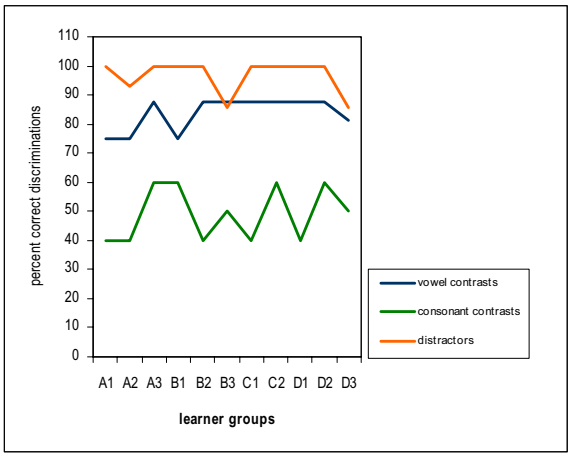
The results presented here are symptomatic of how the variables of age of FL learning, exposure to the FL, dominant L1(s), and gender had (or did not have) an effect on the discrimination of English sounds by several FL learner groups. However, they do not show whether a particular type of sound – vowel or consonant – was either easy or difficult to some extent to perceive by all subjects. The supposition that there was a certain degree of difficulty in discriminating sound contrasts lies in the finding that none of the 11 learner groups obtained a total correct discrimination score approximating 19 or 20 (out of 20). Moreover, within the perceptual task, correct discrimination scores varied greatly depending on whether groups discriminated minimal pairs or distractors. As seen in Figure 5.6, learners identified distractors at higher rates (range = 85.71% – 100%) than minimal pairs containing both vowel and consonant sound oppositions[70]. Minimal pairs focusing on vowel contrasts were also discriminated more often as correct in comparison to minimal pairs containing consonant contrasts (range = 75% – 87.50% for vowels, and 40% – 60% for consonants). To find out if the differences in correct scores obtained for each type of sound contrast and distractors were significant, those scores were submitted

[70] The only exception was group B3, whose correct discrimination rate for vowel contrasts was slightly higher than that of distractors (87.50% vs. 85.71%).

to a Friedman test[71]. Results showed that the differences observed were indeed significant (χ^2 26.968, *df* 2, $p < .05$).

In addition to the finding of significant differences in the discrimination of the three types of pairs examined in the perceptual task, a Cochran *Q* test revealed that the 20 pairs in the auditory discrimination task in comparison to each other were not always equally easy or difficult to discriminate by any of the subject groups (χ^2 1999.957, *df* 19, $p < .05$). Thus, while there seemed to be uniformity among groups as to what type of contrast was discriminated correctly at higher rates – in descending order: distractors, vowels, and consonants – subjects differed in the number of pairs and the specific contrasts that they found easy to discriminate in relation to other contrasts included in the task[72].

Figure 5.6. Median percent correct discriminations of vowel contrasts, consonant contrasts, and distractors.



If groups did not perceive as 100% correct the same number and type of pairs, a similar pattern emerged for pairs that had a certain degree of difficulty. One study that has provided a description of the different degrees of difficulty encountered by learners in

[71] The number of pairs included to test for each type of contrast did not match: 8 pairs for vowels, 5 for consonants, and 7 for distractors. As a result, original discrimination scores (raw scores) were previously converted to *z* scores, so that cross-comparisons could be made.

[72] According to this test, a mean of 1.00 for a given pair meant that it was successfully discriminated all the time, whereas means lower than 1.00 meant that a group did not discriminate a specific pair correctly 100% of the time.

the discrimination of English sounds is García Lecumberri (1999). In that study the perception of English vowel sounds by Spanish NSs was examined, resulting in a classification of sounds according to their perceptual difficulty. Thus, if the correct identification rate of specific sounds was above 80%, the sounds in question were considered to have little difficulty for their accurate discrimination. If the correct identification rate was between 65% and 80%, sounds were thought to present some difficulty. And, finally, if the rate was below 65%, English vowel sounds were considered to present a great deal of difficulty for Spanish-speaking subjects to discern correctly.

The cut-off point of 80% in the García Lecumberri study (1999) rests on IL phonology studies, whereby it has been established that if learners produce and/or perceive an L2 sound accurately in 80% of instances, that L2 sound is considered to have been acquired (see, for example, Carlisle, 1998; see also comments/observations by Riney & Flege, 1998). It should be noted that while in those studies subjects perceived or produced the same sound or sound contrasts a number of times, in the AX task subjects performed for this study, a specific sound contrast was presented just once. And in those cases where it was presented twice, the sound opposition appeared in two different phonetic environments (e.g. /æ/-/ɛ/ contrast was presented in a voiced plosive context, [bVd], and in a voiced nasal context, [mVn]).

Based on García Lecumberri's study and criteria, a classification of different degrees of difficulty in the discrimination of English sounds included in the AX task (together with the results of the Cochran *Q* test as well as the frequencies computed) may be outlined as follows[73]:

(a) Sound contrasts that present no difficulty: 100% correct discrimination
(b) Sound contrasts that present little difficulty: correct discrimination rate higher than 80% (and just below 100%)
(c) Sound contrasts that present some difficulty: correct discrimination rate between 65% and 80%
(d) Sound contrasts that present a great deal of difficulty: correct discrimination rate below 65%

Table 5.3 shows the distribution/classification of the sound contrasts in the AX task arranged in these four categories by learner groups.

[73] The results obtained for English sound contrasts in this study should be considered "tentative", given the limited number of presentations for a particular sound contrast.

Table 5.3. Classification of sound contrasts according to their perceptual difficulty by the learner groups of the study.

(a)

Degree of perceived difficulty in discerning sound contrasts	**8-year-old beginners** (group A) *sound contrasts*		**11-year-old beginners** (group B) *sound contrasts*	
No difficulty: 100% correct discrimination	/æ/-/ɛ/[8,12]	A3	/æ/-/ɛ/[8,12] /ɒ/-/ʌ/[7] /ɒ/-/ɑ/	B1[(12)], B2, B3 B2, B3 B2
Little difficulty: 81%–99% correct discrimination	/æ/-/ɛ/[8,12] /ɒ/-/ʌ/[7,14] /ɒ/-/ɑ/ /ɪ/-/ɛ/ /ɪ/-/i/[11] /b/-/v/	A1, A2 A1, A2, A3 A2, A3 A2, A3 A3 A3	/æ/-/ɛ/[8] /ɒ/-/ʌ/[7,14] /ɒ/-/ɑ/ /ɪ/-/ɛ/ /b/-/v/	B1 B1,B2[(14)],B3[(14)] B1, B3 B1, B2, B3 B1, B3
Some difficulty: 65%–80% correct discrimination	/ɪ/-/ɛ/ /ɒ/-/ɑ/ /b/-/v/ /ʒ/-/dʒ/	A1 A1 A1 A2, A3	/ɪ/-/i/[11] /b/-/v/ /ʒ/-/dʒ/	B3 B2 B1, B2, B3
High difficulty: below 65% correct discrimination	/i/-/ɪ/[1,11] /ʒ/-/dʒ/ /b/-/v/ /t/-/d/ /b/-/p/ /s/-/z/	A1, A2, A3[(1)] A1 A2 A1, A2, A3 A1, A2, A3 A1, A2, A3	/i/-/ɪ/[1,11] /t/-/d/ /b/-/p/ /s/-/z/	B1, B2, B3[(1)] B1, B2, B3 B1, B2, B3 B1, B2, B3

[1]pair # 1 *seat-sit,* [7]pair # 7 *gone-gun,* [8]pair # 8 *man-men,* [11]pair # 11 *still-steal,* [12]pair # 12 *bad-bed,* [14]pair # 14 *cop-cup*

(b)

Degree of perceived difficulty in discerning sound contrasts	**14-year-old beginners** (group C) *sound contrasts*		**adult beginners** (group D) *sound contrasts*	
No difficulty: 100% correct discrimination	/æ/-/ɛ/[8,12] /ɒ/-/ɑ/ /ɒ/-/ʌ/[14]	C1[(12)], C2 C1, C2 C2	/æ/-/ɛ/[8,12] /ɒ/-/ɑ/ /ɒ/-/ʌ/[7,14] /ɪ/-/ɛ/	D1, D2, D3 D3 D2, D3 D2, D3
Little difficulty: 81%–99% correct discrimination	/ɪ/-/i/[11] /ɪ/-/ɛ/ /æ/-/ɛ/[8] /ɒ/-/ʌ/[7,14] /b/-/v/	C1, C2 C1, C2 C1 C1, C2[(14)] C2	/ɪ/-/i/[11] /ɪ/-/ɛ/ /ɒ/-/ɑ/ /ɒ/-/ʌ/[7,14] /b/-/v/ /ʒ/-/dʒ/	D2 D1 D1, D2 D1 D1 D2
Some difficulty: 65%–80% correct discrimination	/b/-/v/ /ʒ/-/dʒ/	C1 C1, C2	/ɪ/-/i/[11] /s/-/z/ /b/-/v/ /ʒ/-/dʒ/	D1 D2, D3 D2 D1, D3
High difficulty: below 65% correct discrimination	/i/-/ɪ/[1] /t/-/d/ /b/-/p/ /s/-/z/	C1, C2 C1, C2 C1, C2 C1, C2	/i/-/ɪ/[1,11] /b/-/v/ /t/-/d/ /b/-/p/ /s/-/z/	D1[(1)], D2[(1)], D3 D3 D1, D2, D3 D1, D2, D3 D1

[1]pair # 1 *seat-sit,* [7]pair # 7 *gone-gun,* [8]pair # 8 *man-men,* [11]pair # 11 *still-steal,* [12]pair # 12 *bad-bed,* [14]pair # 14 *cop-cup*

Due to the statistical significant differences found in the discrimination of minimal pairs vs. distractors, the possible effects of the variables of the study on the correct discrimination of each type of pair (vowel contrasts, consonant contrasts, and distractors) were further explored. Besides, the areas of vowel and consonant perception have normally been examined separately in L2 phonological acquisition research (for vowel perception, see, for example, Cebrian, 2002c; Flege, 1991a; Flege et al., 1994, 1997, 1999; Fox et al., 1995; García Lecumberri & Gallardo, 2003; and Rallo, 2005; for consonant perception, see, e.g., Bradlow et al., 1997; Flege & Eefting, 1987; Guion et al., 2000; and MacKay et al., 1997, 2001; cf. Snow & Hoefnagel-Höhle, 1978/1982). So, additional Kruskal-Wallis analyses (or Mann-Whitney *U* tests for two-level factors) were carried out on the total correct responses to all vowel contrasts, consonant contrasts, and distractors as separate dependent variables; and age of FL learning, exposure to the FL, dominant L1(s), and gender as factors. When these analyses yielded significant results, Mann-Whitney *U* tests were performed to locate between which groups the differences were significant. For all these analyses a rather conservative alpha level was used, so that the experiment-wise error was held constant at .05. In this case, the alpha level was set at .016 to maintain an experiment-wise error of .05 (.016 x 3 types of pair – vowel contrasts, consonant contrasts, and distractors).

5.1.5.1. Effects of onset age of FL learning, exposure, dominant L1(s), and gender on the discrimination of vowel contrasts

As far as vowel contrasts are concerned, onset age of FL learning as a factor was significant for A1, B1, C1, and D1 (χ^2 21.245, *df* 3, adjusted $p < .05$) (Figure 5.7a), as well as for A2, B2, and D2 (χ^2 22.020, *df* 2, adjusted $p < .05$) (Figure 5.7b). Adult learners (D1) discriminated vowel sound oppositions significantly better than 8- year-old beginners at Time 1 (*U* 314.5, *Z* –4.228, adjusted $p < .05$). Mann-Whitney *U* tests also revealed that younger learners (A1) with a mean age of 10.93 years discriminated vowel contrasts significantly more poorly than 14-year-old starters (C1), who were 16.07 years old on average at Time 1 (*U* 147, *Z* – 3.362, adjusted $p < .05$). With 200 hours of instruction in the TL, the remaining comparisons between age groups were nonsignificant. Nevertheless, two comparisons approached significance: one looking at

the differences in scores between A1 and B1 (*U* 286, *Z* –1.972, *p* = .049[74], adjusted *p* > .05); and the other between B1 and D1 (*U* 491.5, *Z* –2.145, *p* = .032, adjusted *p* > .05). In both cases, the older age group included in either comparison discriminated vowel contrasts at higher rates (*M* = 6.07 (B1) vs. 5.34 (A1); 6.73 (D1) vs. 6.07 (B1)).

Age differences were also found between A2 and B2 (*U* 247.5, *Z* –3.805, adjusted *p* < .05), and A2 and D2 (*U* 49.5, *Z* –3.597, adjusted *p* < .05). In both cases, A2 obtained significantly lower correct discrimination scores on vowel contrasts (*M* = 5.61 vs. 6.62 (B2) and 7.10 (D2)). No other two-group comparison reached significance at Time 2.

Finally, age of learning was not a significant factor on the scores obtained by A3 and B3 (*U* 520.5, *Z* –.271, adjusted *p* > .05) (Figure 5.7c). In fact, those two age groups discriminated vowel contrasts almost identically (*M* = 6.56 and 6.53 for A3 and B3, respectively). Last, Table 5.4 provides a summary of all of the group comparisons conducted on the correct scores for vowel contrasts as the dependent variable and onset age of FL learning as a factor.

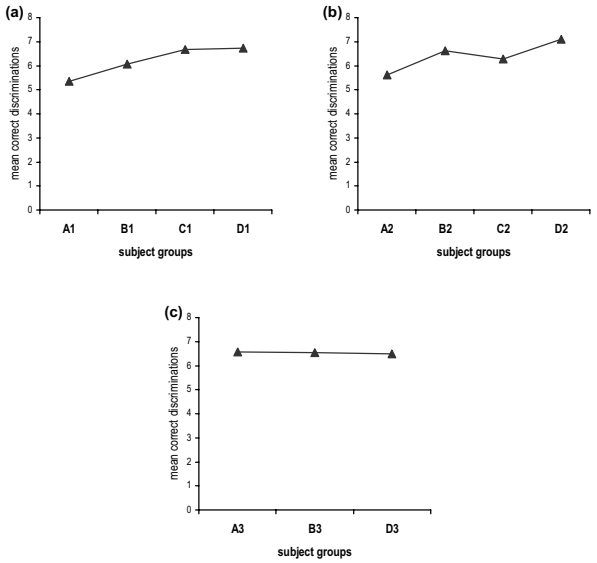
Table 5.4. Summary of comparisons carried out on the discrimination scores for vowel contrasts with AOL as a factor. Significant comparisons are marked with * (*p* < .05), while nonsignificant group comparisons are displayed as n.s. (*p* > .05). *P* values of marginally significant results (.05 < *p* < .10) are stated as *p* = .exact significance value. Each group's mean correct discrimination scores for vowel contrasts appear in parentheses.

Factor: ***onset age of FL learning***
D.V.: correct discrimination scores for vowel contrasts
A1 (5.34) – B1 (6.07) – C1 (6.68) – D1 (6.73)*
A1 < B1 *p* = .049
A1 < C1*
A1 < D1*
B1 – C1 n.s.
B1 < D1 *p* = .032
C1 – D1 n.s.
A2 (5.61) – B2 (6.62) – D2 (7.10)*
A2 < B2*
A2 < D2*
B2 < D2*
A3 (6.56) – B3 (6.53) n.s.

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .016. Thus, * marks a significant result with adjusted *p* < .05. And marginally significant results should be compared to the alpha level of .016 as the resulting adjusted *p* < .05.

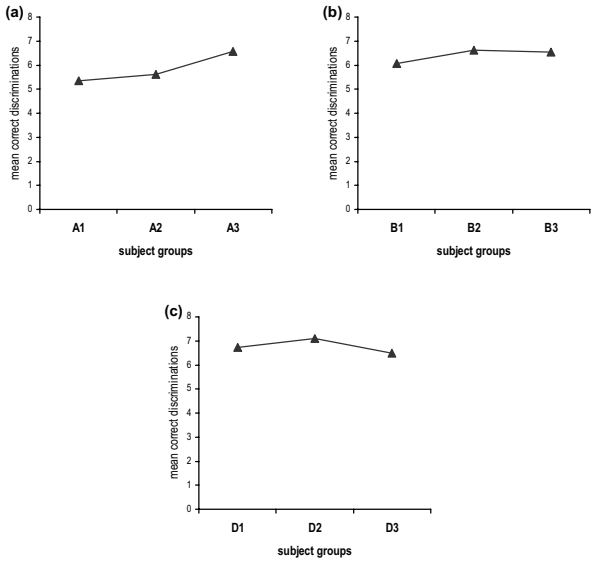
[74] "*p* =" states the raw *p*-value. Therefore, all raw *p*-values reported in 5.1.5.1, 5.1.5.2, and 5.1.5.3 should be interpreted in the light of the new alpha level set – .016 – in order for a result to be considered significant. In this case, *p* = .049 does not mark a significant result, as it exceeds the alpha level of .016.

Figure 5.7. Mean correct discrimination scores for vowel contrasts. Factor: age of onset of FL learning. Results for C2 and D3 are indicative only.



Exposure to the FL yielded significant results in subjects who started to learn the FL at 8 years of age (χ^2 17.200, *df* 2, adjusted $p < .05$) (Figure 5.8a). More precisely, a significant improvement in the discrimination of vowel contrasts was observed along with an increase in exposure for group A3 (726 hours) in comparison to A1 (200 hours) and A2 (416 hours) (M = 6.56 vs. 5.34 and 5.61) (U 179.5, Z –3.637, adjusted $p < .05$; and U 236.5, Z –3.652, adjusted $p < .05$, respectively). But from 200 hours to 416 hours of exposure, learners did not differ in their scores significantly (U 487.5, Z –.474, adjusted $p > .05$), though A2 (M = 5.61) obtained slightly higher correct discrimination scores than A1 (M = 5.34).

Figure 5.8. Mean correct discriminations for vowel contrasts. Factor: exposure to the FL. Results for D3 are indicative only.



Subjects who were first exposed to English at 11 years old did not discriminate vowel contrasts significantly better as experience in the FL increased (χ^2 2.598, *df* 2, adjusted $p > .05$). Moreover, mean correct discrimination scores varied little, as exposure increased. However, there was a slight improvement in subjects' discernment of vowel contrasts, as their amount of formal instruction in English increased: 6.07, 6.62, and 6.53 for B1, B2, and B3, respectively (Figure 5.8b).

At a descriptive level, an increase in experience in 14-year-old beginners did not lead to higher discrimination scores for vowel contrasts (M = 6.68 and 6.29 for C1 and C2). But it should be noted once more that C2 consisted of a small sample of subjects.

For adults, an increase in experience (e.g. from 200 hours to 416 hours) did result in a better perception of vowel contrasts (M = 6.73 and 7.10 for D1 and D2, respectively), but the differences were not large enough to be significant (U 198.5, Z –.987, adjusted $p > .05$) (Figure 5.8c).

Table 5.5 below summarises all the comparisons carried out on the correct discrimination scores for vowel contrasts with exposure to the FL as a factor.

Table 5.5. Summary of comparisons carried out on the discrimination scores for vowel contrasts with exposure to the FL as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). Each group's mean correct discrimination scores for vowel contrasts appear in parentheses.

Factor: ***exposure***   *D.V.:* correct discrimination scores for vowel contrasts
A1 (5.34) – A2 (5.61) – A3 (6.56)*   A1 – A2 n.s.   A1 < A3*   A2 < A3*   B1 (6.07) – B2 (6.62) – B3 (6.53) n.s.   D1 (6.73) – D2 (7.10) n.s.

Note: for a result to be significant at the .05 level, the significance level should be <= .016. Thus, * marks a significant result with adjusted $p < .05$.

As for the effect of dominant L1(s) on the subjects' discrimination of vowel contrasts, no language subgroup in any of the subject groups differed significantly from the other two language subgroups (adjusted $p > .05$) (see Table 5.6 for a summary of comparisons). Furthermore, being a Catalan or Spanish dominant speaker, or Catalan/Spanish balanced bilingual did not entail a consistent, better discernment of English vowel contrasts by any of the language subgroups, as shown in Figure 5.9 below.

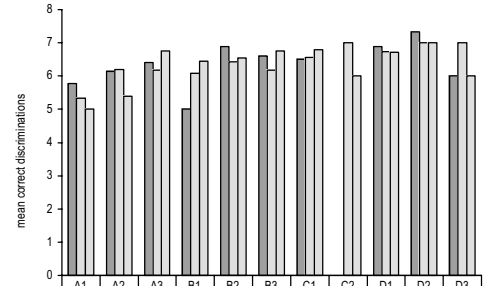
Table 5.6. Summary of comparisons carried out on the discrimination scores for vowel contrasts with dominant L1(s) as a factor. Nonsignificant group comparisons are displayed as n.s. ($p > .05$). Each group's mean correct discrimination scores for consonant contrasts appear in parentheses.

Factor: ***dominant L1(s)***   *D.V.:* correct discrimination scores for vowel contrasts
A1: Cat (5.78) – Sp (5.33) – C/S (5.00) n.s.   A2: Cat (6.14) – Sp (6.20) – C/S (5.39) n.s.   A3: Cat (6.40) – Sp (6.17) – C/S (6.75) n.s.   B1: Cat (5.00) – Sp (6.08) – C/S (6.45) n.s.   B2: Cat (6.89) – Sp (6.43) – C/S (6.54) n.s.   B3: Cat (6.60) – Sp (6.64) – C/S (6.42) n.s.   C1: Cat (6.50) – Sp (6.57) – C/S (6.80) n.s.   D1: Cat (6.88) – Sp (6.73) – C/S (6.72) n.s.   D2: Cat (7.33) – Sp (7.00) – C/S (7.00) n.s.

Cat: Catalan dominant speakers, Sp: Spanish dominant speakers,
C/S: Catalan/Spanish balanced bilinguals

Note: for a result to be significant at the .05 level, the significance level should be <= .016.

Figure 5.9. Mean correct discriminations for vowel contrasts by Catalan dominant speakers, Spanish dominant speakers, and Catalan/Spanish balanced bilinguals in each learner group. Results for C2 and D3 are indicative only.

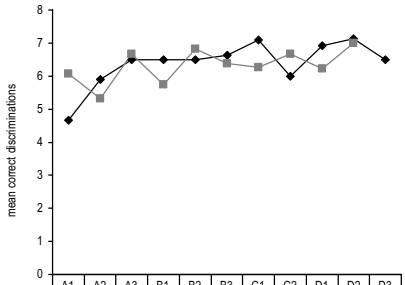


	A1	A2	A3	B1	B2	B3	C1	C2	D1	D2	D3
Catalan	5.78	6.14	6.4	5	6.89	6.6	6.5		6.88	7.33	6
Spanish	5.33	6.2	6.17	6.08	6.43	6.17	6.57	7	6.73	7	7
Cat/Sp	5	5.39	6.75	6.45	6.54	6.75	6.8	6	6.72	7	6

subject groups

Last, as seen in Figure 5.10, male and female subjects in each learner group discriminated English vowel sounds on a similar basis. Also, no common pattern as to any of the gender subgroups' higher discrimination scores could be outlined. In other words, there was a lot of variability between male and female subjects' correct discrimination scores. The largest difference in scores found was between male and female speakers in A1 (M = 6.07 and 4.67, respectively), which only approached significance (U 55.5, Z –2.243, p = .025, adjusted p > .05). No other difference approached or reached significance (adjusted p > .05) (see Table 5.7).

Figure 5.10. Mean correct discriminations for vowel contrasts by male and female subjects in each learner group. Results for C2 and D3 are indicative only.



	A1	A2	A3	B1	B2	B3	C1	C2	D1	D2	D3
female	4.67	5.89	6.5	6.5	6.5	6.63	7.09	6	6.92	7.13	6.5
male	6.07	5.33	6.67	5.75	6.82	6.38	6.27	6.67	6.23	7	

subject groups

Table 5.7. Summary of comparisons carried out on the discrimination scores for vowel contrasts with gender as a factor. Nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each group's mean correct discrimination scores for vowel contrasts appear in parentheses.

Factor: ***gender***
D.V.: correct discrimination scores for vowel contrasts
A1: Female (4.67) – Male (6.07) ***p*** = .025
A2: Female (5.89) – Male (5.33) n.s.
A3: Female (6.50) – Male (6.67) n.s.
B1: Female (6.50) – Male (5.75) n.s.
B2: Female (6.50) – Male (6.82) n.s.
B3: Female (6.63) – Male (6.38) n.s.
C1: Female (7.09) – Male (6.27) n.s.
D1: Female (6.92) – Male (6.23) n.s.
D2: Female (7.13) – Male (7.00) n.s.

Note: for a result to be significant at the .05 level, the significance level should be <= .016. Marginally significant results should be compared to the alpha level of .016 as the resulting adjusted $p < .05$.

5.1.5.2. Effects of onset age of FL learning, exposure, dominant L1(s), and gender on the discrimination of consonant contrasts

With reference to consonant contrasts, onset age of English learning did not have a significant effect on the correct discrimination scores obtained by A1, B1, C1, and D1 (χ^2 9.642, *df* 3, $p = .022$, adjusted $p > .05$), though it approached significance. Nor was the effect of AOL significant on the scores obtained by A2, B2, and D2 (χ^2 2.494, *df* 2, adjusted $p > .05$), and those of A3 and B3 (*U* 467, *Z* –.995, adjusted $p > .05$) (see Table 5.8). In fact, only half of the consonant contrasts presented in the AX task were discerned correctly (on average, 2.5 out of 5) among all age groups (see Figure 5.11).

Figure 5.11. Mean correct discrimination scores for consonant contrasts. Factor: age of onset of FL learning. Results for C2 and D3 are indicative only.

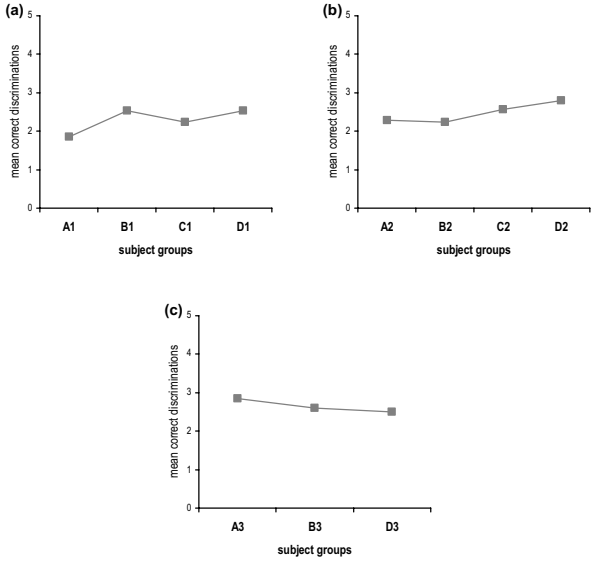


Table 5.8. Summary of comparisons carried out on the discrimination scores for consonant contrasts with AOL as a factor. Nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each group's mean correct discrimination scores for consonant contrasts appear in parentheses.

Factor: ***onset age of FL learning***   *D.V.:* correct discrimination scores for consonant contrasts
A1 (1.86) – B1 (2.54) – C1 (2.23) – D1 (2.53) ***p*** **=** .022
A2 (2.28) – B2 (2.24) – D2 (2.80) n.s.
A3 (2.85) – B3 (2.60) n.s.

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .016. Marginally significant results should be compared to the alpha level of .016 as the resulting adjusted $p < .05$.

When amount of exposure was examined, differences in scores between A1, A2, and A3 (M = 1.86, 2.28, and 2.85) reached significance (χ^2 10.732, df 2, p = .005, adjusted $p < .05$) (Figure 5.12a); hence, paralleling the significant effect of exposure on

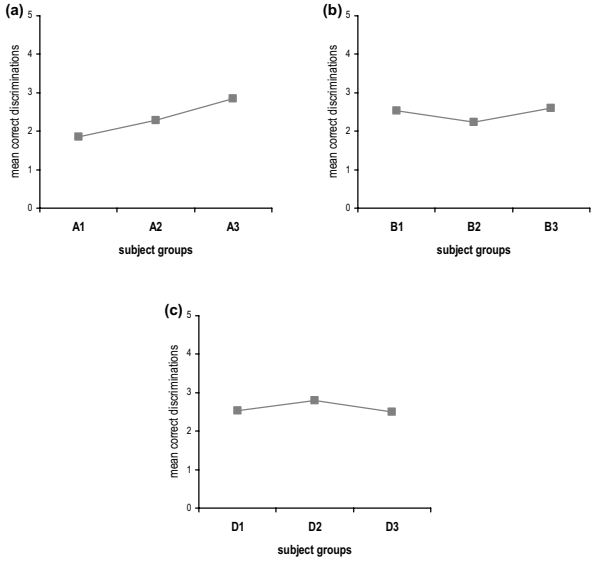
the same 8-year-old beginners in the discrimination of vowel contrasts. That is, an increase in exposure led to early beginners' better discrimination of consonant contrasts, A3's discrimination scores being significantly higher than those of A1 (*U* 199.5, *Z* –3.270, *p* = .001, adjusted *p* < .05) and close to being significant in relation to A2's scores (*U* 355, *Z* –1.893, *p* = .058, adjusted *p* > .05). In the case of the between-group comparison of A1 and A2, the differences in scores were not statistically significant, according to a Mann-Whitney *U* test (*U* 406.5, *Z* –1.578, *p* = .114, adjusted *p* > .05) (Figure 5.12b,c). A similar effect of exposure (i.e. higher correct discrimination scores along with an increase in experience in the TL) was observed in 14-year-old beginners – although not analysed statistically for the reasons outlined above – (*M* = 2.23 and 2.57 for C1 and C2, respectively). The remaining group comparisons with exposure as a factor were nonsignificant: B1, B2, and B3 (*M* = 2.54, 2.24, and 2.60) (χ^2 2.679, *df* 2, adjusted *p* > .05); and D1 and D2 (*M* = 2.53 and 2.80) (*U* 203, *Z* –.897, adjusted *p* > .05). In general terms, an increase in exposure resulted in 11-year-old beginners' and adult learners' higher rates of correct discrimination scores of consonant contrasts. However, it should be noted that 11-year-old beginners' discrimination of consonant contrasts became somewhat poorer halfway through their learning (B2), though not in a significant way, as stated above. Table 5.9 below provides a summary of all of the group comparisons carried out on the correct discrimination scores for consonant contrasts with exposure to the FL as a factor.

Table 5.9. Summary of comparisons carried out on the discrimination scores for consonant contrasts with exposure to the FL as a factor. Significant comparisons are marked with * (*p* < .05), while nonsignificant group comparisons are displayed as n.s. (*p* > .05). *P* values of marginally significant results (.05 < *p* < .10) are stated as *p* = .exact significance value. Each group's mean correct discrimination scores for consonant contrasts appear in parentheses.

Factor: ***exposure***   *D.V.:* correct discrimination scores for consonant contrasts
A1 (1.86) – A2 (2.28) – A3 (2.85)*   A1 – A2 n.s.   A1 < A3*   A2 < A3 ***p*** = .058
B1 (2.54) – B2 (2.24) – B3 (2.60) n.s.
D1 (2.53) – D2 (2.80) n.s.

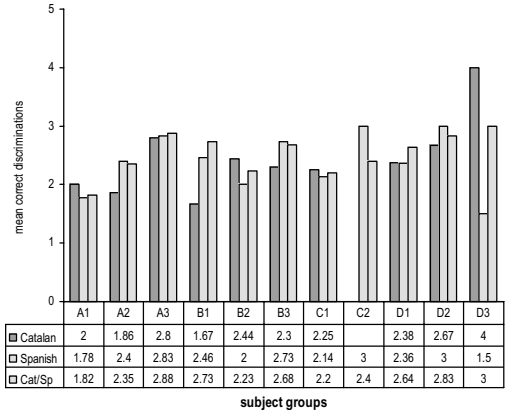
Note: for a result to be significant at the .05 level, the significance level should be <= .016. Thus, * marks a significant result with adjusted *p* < .05. And marginally significant results should be compared to the alpha level of .016 as the resulting adjusted *p* < .05.

Figure 5.12. Mean correct discriminations for consonant contrasts. Factor: exposure to the FL. Results for D3 are indicative only.



As seen in Figure 5.13 below, the three language subgroups in each learner group obtained differing discrimination scores for consonant contrasts. Other than the finding of Spanish/Catalan balanced bilinguals with 416 hours of exposure (A2, B2, and D2) who discriminated consonant contrasts at correct rates between those of learners with 200 hours and 726 hours of formal instruction in their respective groups, no definite pattern could be outlined for a specific language subgroup's advantage (or disadvantage) over the other two language subgroups across learners. The variability found among the three language subgroups in each subject group did not result in any significant differences based on Kruskal-Wallis tests (adjusted $p > .05$) (see also Table 5.10).

Figure 5.13. Mean correct discriminations for consonant contrasts by Catalan dominant speakers, Spanish dominant speakers, and Catalan/Spanish balanced bilinguals in each learner group. Results for C2 and D3 are indicative only.



	A1	A2	A3	B1	B2	B3	C1	C2	D1	D2	D3
Catalan	2	1.86	2.8	1.67	2.44	2.3	2.25		2.38	2.67	4
Spanish	1.78	2.4	2.83	2.46	2	2.73	2.14	3	2.36	3	1.5
Cat/Sp	1.82	2.35	2.88	2.73	2.23	2.68	2.2	2.4	2.64	2.83	3

Table 5.10. Summary of comparisons carried out on the discrimination scores for consonant contrasts with dominant L1(s) as a factor. Nonsignificant group comparisons are displayed as n.s. (p > .05). Each group's mean correct discrimination scores for consonant contrasts appear in parentheses.

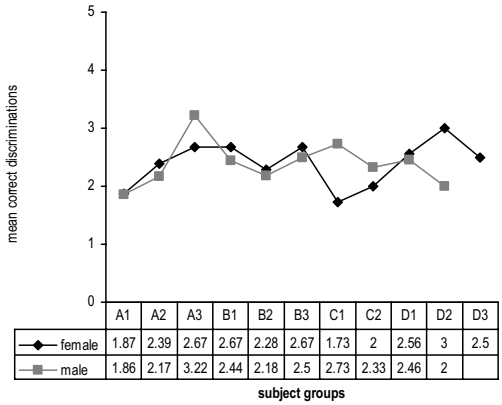
Factor: ***dominant L1(s)*** *D.V.:* correct discrimination scores for consonant contrasts	
A1: Cat (2.00) – Sp (1.78) – C/S (1.82)	n.s.
A2: Cat (1.86) – Sp (2.40) – C/S (2.35)	n.s.
A3: Cat (2.80) – Sp (2.83) – C/S (2.80)	n.s.
B1: Cat (1.67) – Sp (2.46) – C/S (2.73)	n.s.
B2: Cat (2.44) – Sp (2.00) – C/S (2.23)	n.s.
B3: Cat (2.30) – Sp (2.73) – C/S (2.68)	n.s.
C1: Cat (2.25) – Sp (2.14) – C/S (2.20)	n.s.
D1: Cat (2.38) – Sp (2.36) – C/S (2.64)	n.s.
D2: Cat (2.67) – Sp (3.00) – C/S (2.83)	n.s.

Cat: Catalan dominant speakers, Sp: Spanish dominant speakers, C/S: Catalan/Spanish balanced bilinguals

Note: for a result to be significant at the .05 level, the significance level should be <= .016.

As for the variable of gender, there was a tendency for female speakers in 11-year-old and adult starter groups to discriminate consonant sound oppositions at higher correct rates than male subjects matched for AOL and exposure. However, this tendency did not generalise over 8- and 14-year-old beginners, as shown in Figure 5.14. On two occasions the scores between male and female speakers were one score different, namely C1 and D2 (the maximum number of correct scores was 5). Mann-Whitney *U* tests showed that those differences were only close to being significant (for C1: *U* 35.5, *Z* –1.703, *p* = .089, adjusted *p* > .05; and for D2: *U* 2, *Z* –1.677, *p* = .094, adjusted *p* > .05). Mann-Whitney *U* tests further revealed that the differences in scores for consonant contrasts between male and female subjects in the remaining learner groups were not statistically significant (Table 5.11).

Figure 5.14. Mean correct discriminations for consonant contrasts by male and female subjects in each learner group. Results for C2 and D3 are indicative only.



	A1	A2	A3	B1	B2	B3	C1	C2	D1	D2	D3
female	1.87	2.39	2.67	2.67	2.28	2.67	1.73	2	2.56	3	2.5
male	1.86	2.17	3.22	2.44	2.18	2.5	2.73	2.33	2.46	2	

Table 5.11. Summary of comparisons carried out on the discrimination scores for consonant contrasts with gender as a factor. Nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each group's mean correct discrimination scores for consonant contrasts appear in parentheses.

Factor: ***gender***   *D.V.:* correct discrimination scores for consonant contrasts
A1: Female (1.87) – Male (1.86) n.s.
A2: Female (2.39) – Male (2.17) n.s.
A3: Female (2.67) – Male (3.22) n.s.
B1: Female (2.67) – Male (2.44) n.s.
B2: Female (2.28) – Male (2.18) n.s.
B3: Female (2.67) – Male (2.50) n.s.
C1: Female (1.73) – Male (2.73) ***p*** = .089
D1: Female (2.56) – Male (2.46) n.s.
D2: Female (3.00) – Male (2.00) ***p*** = .094

Note: for a result to be significant at the .05 level, the significance level should be <= .016. Marginally significant results should be compared to the alpha level of .016 as the resulting adjusted $p < .05$.

5.1.5.3. Effects of onset age of FL learning, exposure, dominant L1(s), and gender on the discrimination of distractors

Learner groups discriminated distractors correctly at high rates (range = 80%–100%). In addition, a Cochran *Q* test revealed that overall all groups discerned the various distractors on a similar basis (Q= 4 – 12.293, df 6, $p > .05$), except for A1 and B3 (Q 15, df 6, $p < .05$; and Q 14.727, df 6, $p < .05$). In both cases, the pair of distractors that was discriminated at poorer rates was pair #2 (see Appendix A): 86% and 79% for A1 and B3, respectively.

As shown in Figure 5.15, the four age groups obtained similar scores. Therefore, Kruskal-Wallis tests with onset age of FL learning as a factor did not yield any significant differences in the scores obtained by the various age groups (adjusted $p > .05$) (see Table 5.12).

Figure 5.15. Mean correct discrimination scores for distractors. Factor: age of onset of FL learning. Results for C2 and D3 are indicative only.

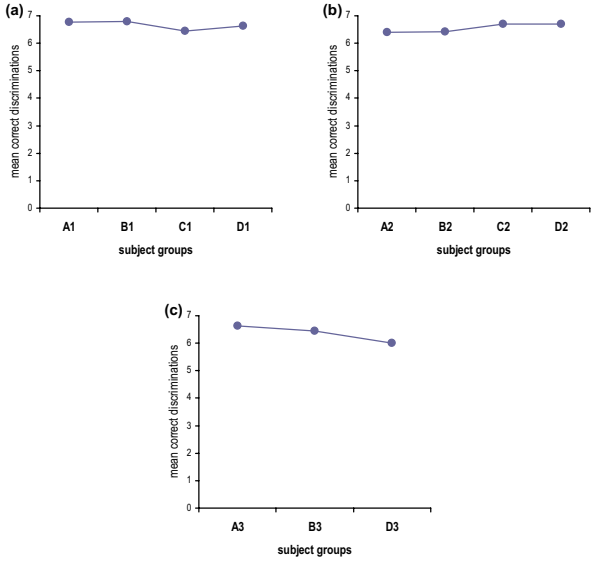


Table 5.12. Summary of comparisons carried out on the discrimination scores for distractors with AOL as a factor. Nonsignificant group comparisons are displayed as n.s. (p > .05). Each group's mean correct discrimination scores for distractors appear in parentheses.

Factor: ***onset age of FL learning***   *D.V.:* correct discrimination scores for distractors
A1 (6.76) – B1 (6.79) – C1 (6.45) – D1 (6.63) n.s.
A2 (6.39) – B2 (6.41) – D2 (6.70) n.s.
A3 (6.63) – B3 (6.45) n.s.

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .016.

Regardless of their amount of exposure, all groups obtained discrimination scores above 6 (out of a maximum of 7) (see Figure 5.16a,b,c). An increase in the amount of formal instruction in the FL did not lead to a significantly better discrimination of

distractors. Moreover, 8- and 11-year-old starters discriminated distractors at lower rates when they had 416 hours of instruction (M = 6.39 for A2 and 6.41 for B2) in comparison to learners with 200 hours (M = 6.76 for A1 and 6.79 for B1) and 726 hours of exposure to English (M = 6.63 for A3 and 6.45 for B3). In fact, those age groups obtained higher scores when they had the least amount of exposure according to the research design. These differences in scores only approached significance for A1, A2, and A3 (χ^2 5.394, *df* 2, p = .067, adjusted p > .05). By contrast, the differences in scores were indeed significant for B1, B2, and B3 (χ^2 9.915, *df* 2, p = .007, adjusted p < .05). As stated above, the group with the least amount of exposure – B1 – obtained significantly higher correct discrimination scores than B2 and B3 (U 271.5, Z –2.639, p = .008, adjusted p < .05; and U 355, Z –3.035, p = .002, adjusted p < .05). In the case of 14-year-old starters, there was an improvement in the discrimination of distractors, as exposure increased (M = 6.45 and 6.71 for C1 and C2, respectively). The same effect, though not significant, was observed for adult learners (M = 6.63 and 6.70 for D1 and D2, respectively) (U 220, Z –.627, adjusted p > .05) (see Table 5.13).

Figure 5.16. Mean correct discriminations for distractors. Factor: exposure to the FL. Results for D3 are indicative only.

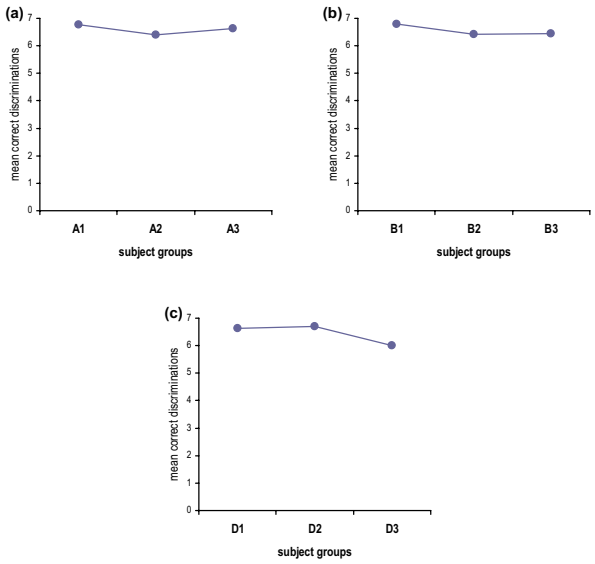


Table 5.13. Summary of comparisons carried out on the discrimination scores for distractors with exposure to the FL as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each group's mean correct discrimination scores for distractors appear in parentheses.

Factor: ***exposure***
D.V.: correct discrimination scores for distractors
A1 (6.76) – A2 (6.39) – A3 (6.63) ***p*** = .067
B1 (6.79) – B2 (6.41) – B3 (6.45)* B1 > B2* B1 > B3* B2 – B3 n.s.
D1 (6.63) – D2 (6.70) n.s.

Note: for a result to be significant at the .05 level, the significance level should be <= .016. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .016 as the resulting adjusted $p < .05$.

A final comment on distractor discrimination has to do with the variables of dominant L1(s) and gender.

Subjects' dominant L1(s) did not have any effect on the discrimination of distractors. As was the case of the discrimination of vowel and consonant contrasts, there was no pattern as to what language subgroup consistently obtained higher or lower discrimination scores across all learner groups. In addition, Krukal-Wallis tests showed that there were no instances of significant differences in scores between the three language subgroups in any of the subject groups (adjusted $p > .05$). Only on one occasion were the differences in scores close to being significant, namely in D2 (χ^2 5.185, *df* 2, p = .075, adjusted $p > .05$)[75].

Last, Mann-Whitney *U* tests did not reveal any significant differences in the discrimination scores for distractors between male and female subjects (adjusted $p > .05$). In spite of this, there was a tendency for female subjects to obtain higher discrimination scores than male subjects matched for AOL and exposure, especially for 8 and 14-year-old beginners with varying degrees of exposure.

[75] For results obtained for both the variable of dominant L1(s) and that of gender, figures and tables are not presented, due to the lack of significance of the comparisons carried out, and to the fact that all learners discriminated distractors at very high correct rates, as reported above.

5.2. Imitation task

To examine learners' productions of FL sounds, two studies were undertaken after the pilot study reported in the Analyses section (4.4.2): Study 1, which looked at data collected until the spring of 1999, and Study 2, which looked at data from all learner groups on the research project.

The purpose of Study 1 was to determine the effect(s) of the research variables of onset age of FL learning, exposure to the FL, dominant L1(s), and gender on the global foreign accent ratings obtained for six words in the imitation task, in addition to six segments (three vowels and three consonants) that the six words contained (one segment per word). Study 2 further investigated the effect(s) of the research variables mentioned above by focusing on the study of seven English vowel sounds. In that study, the English vowels produced by research subjects were rated for degree of foreign accent, and later transcribed (identified) in a forced-choice identification task. For the most part, the process of digitisation in Studies 1 and 2 (or Experiments 1 and 2) was carried out at the *Laboratori de Fonètica* of the *Universitat de Barcelona*. Both studies were set up and conducted at the Linguistics Research Laboratory of the University of Ottawa.

Prior to the detailed description of each of Study 1 and Study 2, the following methodological issues considered in the design and implementation of both experiments are worth mentioning. First, listening conditions were controlled for. That is, the presentation of stimuli under investigation was administered to judges (or listeners) in the same number of sessions. Moreover, each session focused on a given segment or word at a time. Judges' records were automatically saved on a computer as they performed the rating task. The design of both studies did not allow for judges' change in ratings once they had assigned a rating to a specific subject's production. However, the design did allow for further listening (as many times as required) before rating. Second, the number of raters (or listeners or judges) was increased: from two in the pilot study to six and seven in Study 1 and Study 2, respectively. And third, a control group of English native speakers was included in both experiments to prevent listeners from making biased judgements of learners' performance as consistently foreign-accented.

5.2.*I.* STUDY 1

As follows, the methodology, analyses, and results obtained for accent ratings on words and specific segments of Study 1 are presented. Note that the discussion of results will be included in Chapter 6, together with discussion of results on AX task and Study 2.

5.2.1. Method

5.2.1.1. Subjects

Out of the 281 subjects that comprised the sample examined in this dissertation, 211 learners were investigated in this study. The characteristics of both cross-sectional and longitudinal subjects are displayed in Tables 5.14 and 5.15[76], respectively. It should be noted that the criterion adopted to obtain a cross-sectional data matrix was the same as the one followed in the design of the cross-sectional data matrix for the AX task. However, if Table 4.3 (Section 4.1) and Table 5.14 (below) are compared, it can be readily observed that the number of subjects in each table does not match. In the case of group A (8-year-old beginners), more cross-sectional subjects were present at Times 1 and 2 in the production task than in the perceptual task (32 Ss and 43 Ss at T1 and T2 in the production task vs. 29 Ss and 36 Ss at T1 and T2 in the AX task). This difference in the number of subjects is due to the fact that when Study 1 (imitation task) was carried out, subject data for A3 had not yet been collected. Therefore, there were fewer longitudinal Ss across Time 1 and Time 2 than the sample of subjects studied in the AX task. Accordingly, only half of the subjects were randomly removed from Time 1 and kept at Time 2 (as was the case of group B), which resulted in somewhat larger group sizes for A1 and A2 in Experiment 1. As for group B, there was one less subject in B1 in this experiment than in the AX task. The missing learner's performance was not tape-recorded when it came to doing the production task. Finally, with regard to C1 and D1, not all subjects belonging to those groups had performed the phonetic tasks at the time the larger research project of which this study is a part was set up.

[76] As stated above, longitudinal Ss are not examined in this dissertation. However, tables are presented in both Studies 1 and 2 for descriptive purposes. Also, it should be noted that longitudinal Ss' productions were rated together with those of cross-sectional Ss in all sessions.

A control group of 30 native speakers of British English (henceforth, NE 'native English') was also included in this study. English foils[77] kindly agreed to participate in the study as volunteers. None was fluent in either Spanish or Catalan. This group constituted 12.45% out of the total 241 subjects examined in this study.

Table 5.14. Characteristics of participant groups in Study 1 (cross-sectional data). Standard deviations are in parentheses.

Group	N	AOL[a]	Exposure[b]	L1[c]	Gender[d]	Age[e]	Grade[f]
A1	32	8	200	Cat 28.1% Sp 34.4% C/S 37.5%	m 15 f 17	10.90 (.29)	5 *Primaria*
A2	43	8	416	Cat 23.8% Sp 14.3% C/S 61.9%	m 22 f 21	12.97 (.31)	1 E.S.O.
B1	27	11	200	Cat 10% Sp 46.7% C/S 43.3%	m 16 f 11	13.06 (.35)	7 E.G.B.
B2	29	11	416	Cat 11.5% Sp 46.2% C/S 42.3%	m 11 f 18	15.01 (.31)	1 B.U.P.
B3	40	11	726	Cat 25% Sp 27.5% C/S 47.5%	m 16 f 24	17.95 (.29)	C.O.U.
C1	4	14	200	Cat 25% Sp 25% C/S 50%	m 2 f 2	15.91 (.32)	2 B.U.P.
C2	7	14	416	Cat - Sp 28.6% C/S 71.4%	m 3 f 4	18.70 (.86)	C.O.U.
D1	29	18+	200	Cat 25% Sp 16.7% C/S 58.3%	m 7 f 22	25.20 (4.97)	2 E.I.
NE	30	__	__	___	m 7 f 23	19.43 (11.65)	__

[a] Onset age of FL learning (in years)
[b] Number of hours of formal exposure to English
[c] Dominant L1(s) (%): Cat (Catalan), Sp (Spanish), C/S (Catalan and Spanish)
[d] m: male, f: female
[e] Mean chronological age at testing (in years)
[f] Subjects' school grade at testing:
Primaria and E.S.O. (new curriculum)
E.G.B., B.U.P., and C.O.U. (former curriculum)
E.I. (Escuela de Idiomas) – (adult) language schools

[77] The 30 English NSs included in the control group were visitors on the island of Menorca in the summer of 2000.

Table 5.15. Characteristics of longitudinal subjects in Study 1. Standard deviations are in parentheses.

Group	N	AOL[a]	Exposure[b]	L1[c]	Gender[d]	Age[e]	Grade[f]
A1$_{long*}$	15[g]	8	200	Cat 53.8% Sp 15.4% C/S 30.8%	m 8 f 7	10.94 (.28)	5 *Primaria*
A2$_{long}$	15[g]	8	416	Cat 26.7% Sp 13.3% C/S 60.0%	m 8 f 7	12.97 (.31)	1 E.S.O.
B1$_{long}$	4[h]	11	200	Cat - Sp 50% C/S 50%	m 1 f 3	13.00 (.40)	7 E.G.B.
B2$_{long}$	4[h]	11	416	Cat - Sp 50% C/S 50%	m 1 f 3	15.12 (.41)	1 B.U.P.

*long Longitudinal group
[a] Onset age of FL learning (in years)
[b] Number of hours of formal exposure to English
[c] Dominant L1(s) (%): Cat (Catalan), Sp (Spanish), C/S (Catalan and Spanish)
[d] m: male, f: female
[e] Mean chronological age at testing (in years)
[f] Subjects' school grade at testing:
Primaria and E.S.O. (new curriculum)
E.G.B. and B.U.P. (former curriculum)
[g] Distribution of longitudinal subjects (8-year-old beginners): 15 Ss at Time 1 + Time 2
[h] Distribution of longitudinal subjects (11-year-old beginners): 4 Ss at Time 1 + Time 2

5.2.1.2. Task

The production task that the subjects performed consisted of imitating a list of 34 English words as delivered by a taped model voice of British English via tape-recorder (for a detailed description of the production task, see Section 4.2 and Appendix B). It is worth noting again that subjects' imitations of English words present a considerable amount of background noise, as they were tape-recorded on school premises. At first, the recording equipment consisted of SONY tape recorders Models TCM-313, TCM-459, and TCM-939, and, in later collections, of microphones SONY ECM-717 and VIVANCO EM 216, as well.

5.2.1.3. Objective

The objective of Study 1 was to determine whether the research variables – AOL, exposure, dominant L1(s), and gender – had any effect on learners' productions of FL sounds included in the words imitated. In other words, the purpose was to determine

whether early starters would imitate FL sounds in a more native-like fashion than late starters, or vice versa. It also aimed to find out whether an increase in exposure to the FL would result in a more native-like production of English segments. In addition, the possible effects of dominant L1(s) and gender on the production of TL sounds and words were studied.

For that purpose, learners' word and segment productions were evaluated for degree of foreign accent (FA) by native English speakers. In order to overcome the shortcomings mentioned in the pilot study, the number of raters was increased (from 2 to 6), listening conditions were controlled for (see Section 5.2 above and Procedures below), and a control group of English NSs was added to the experiment.

Since a large number of imitations was obtained – a total of 8,194[78] (241 Ss x 34 words) – the scope of research was narrowed down to six words, namely *jam*, *reading*, *red*, *speak*, *this*, and *very*. Moreover, the segments /æ/, /i/, /d/, /s/, /ɪ/, and /v/ in *jam*, *reading*, *red*, *speak*, *this*, and *very*, respectively, were also selected for further study. Some of the segments chosen in the six words were different from those that had been originally considered for examination (see Appendix B). Hence, the vowel segments selected focused on what could be "tentatively" considered new FL sounds for Spanish and Catalan native speakers – /æ/ and /ɪ/ – and similar FL sounds – /i/ (but see studies reported in Section 2.3.2 above). As for consonant segments, one non-occurring FL sound in Spanish and most varieties of Catalan, namely /v/, was examined. Additionally, two existing consonants in both Spanish and Catalan – /s/ and /d/ – were looked at in contexts where they are not produced in either language: /s/ in absolute word-initial position and /d/ in word-final position.

5.2.1.4. Stimulus preparation

A total of 1,451 imitations was obtained[79]. They were digitised with CoolEdit at 22.05 kHz, 16-bit resolution, and then normalised to 75% peak amplitude.

---

[78] As mentioned earlier, in some instances subjects did not produce one word or a couple of words. So, there are a few missing cases in the production task.

[79] The imitations per word were distributed as follows: 239 for *jam* (2 missing cases), 238 for *reading* (3 missing cases), 240 for *red* (1 missing case), 241 for *speak* (no missing cases), 236 for *this* (5 missing cases), and 237 for *very* (4 missing cases).

For the accent rating task, six different random blocks were created for each of the six words and their imitations. Another six different random blocks were also generated for each of the six segments.

The design and implementation of this experiment via computer was possible thanks to the software Windows Stimulus Presentation and Response Collection System (WNSPARCS) (Smith, 1997).

5.2.1.5. Listeners

Six native speakers of General Canadian English (2 male, 4 female) took part in the study as paid listeners or judges. They were students in Linguistics, Social Sciences, or English Literature at the University of Ottawa. Their mean age was 28 years (range = 22 – 35 years) and reported normal hearing. None of the listeners knew any Catalan or had spent time in a Catalan-speaking community. On the other hand, nearly all were familiar with Spanish (4 judges, in total). Their age of first use of Spanish varied from 16–17 to 30 years, and their formal study ranged from two terms at university to three years in high school (judge 6 in the latter case). Only one judge had spent time in a Spanish-speaking community, namely El Salvador for 2.5 weeks (judge 5). As for languages other than those of the subjects examined in this dissertation, all judges knew at least some French. Their age of first exposure ranged from 2–3 to 10 years, although only one judge used French on a daily basis (judge 6), who, in turn, was the only listener to have actually spent time in a monolingual French-speaking community. As other additional languages, judges had either studied or were exposed at home to one of German, Turkish, Croatian, Russian, Ukrainian, Latin, and Japanese. However, only one judge had used the L2 since childhood.

The choice of Canadian English speakers as listeners (in both Studies 1 and 2) was not thought to be to their disadvantage when assessing subjects' productions of words modelled by a British English voice, as they are used to hearing British English through the media and because there are thousands of British-born residents in the city of Ottawa. In addition, they were explicitly exposed to the models that the FL learner was exposed to, and their attention drawn to the British model.

5.2.1.6. Procedure

Each listener participated in a total of 12 sessions. In the first six sessions, judges assigned a FA rating to each of the six words, as produced by the 241 Ss, on a 9-point scale of FA. The use of a 9-point scale was based on recent research findings that suggest that listeners are able to distinguish up to 9 or 11 different degrees of FA (e.g. Southwood & Flege, 1999). In this study, *1* on the scale meant *no FA* (i.e. native-like production)*, 5* was used to indicate a *moderate amount of FA,* and *9* meant *very strong FA*. The remaining six sessions involved assigning a FA rating to a specific segment in each of the six words on the same 9-point scale of FA. At the beginning of all sessions, listeners were read a set of specific instructions[80] (see Appendix C) and were presented with 5 practice items of the taped model voice that the subjects had imitated. One word was presented at a time and in a different random block to each judge through headphones via computer. The written version of words also appeared on the computer screen while judges listened to the audio files. Judges were told that each block contained both NS and NNS productions. They were also told to ignore background noise as much as possible and to use the whole scale. In all instances, they had to assign a rating by clicking with the mouse on one of the nine scale buttons presented on the computer screen. In order to be able to listen to the next item, listeners had to provide a rating on the specific subject's production. The inter-stimulus interval was 1.5 seconds. However, judges could listen to an item as many times as needed. Each rating session lasted about 35–40 minutes. On the whole, listeners did a single session on a single day. In case they did two sessions, there was an hour interval between the first and second sessions. The rating procedure required 72 sessions – about a month – to be completed (6 judges x 6 words x 2 times [word and segment rating]). A total of 17,172 judgements was obtained: 8,586 FA ratings on words and another 8,586 FA ratings on segments.

[80] Instructions were identical for ratings on global FA of words. In the case of segments, listeners were asked to rate the degree to which the specific vowel segment sounded like a NS of English, and not to pay attention to the (possible) presence of a weak or strong accent on other parts of the word. Further, judges were told to consider /d/ devoicing or elision in *red*, as well as vowel insertion in front of /s/ in *speak*, as instances of FA on the segment in question. This instruction was felt necessary because judges might reason that elision or epenthesis of a segment *before* the target segment might not qualify as FA on the segment.

5.2.2. Results

Judges' accent ratings were analysed by means of the statistical package SPSS 11.0 for Windows. The results[81] obtained for FA ratings on words and on segments will be presented separately.

5.2.2.1. FA ratings on words

In the first place, frequencies of the scale points used to rate all words by each judge were computed. As instructed, listeners used the whole scale virtually in all rating sessions. The only exceptions to this were judges 1 and 6 on assessing the word *speak*. But even on these two occasions when judges did not use the 9 points on the scale, they still made use of an extensive range of the scale, namely points 1–8.

Mean FA ratings of each subject's production of the six words as evaluated by each of the six judges were computed, in addition to mean FA ratings on each of the six words averaged over the six judges. These results are displayed in Figures 5.17 and 5.18, respectively. Besides, mean FA ratings for each group according to their dominant L1(s) and gender were computed and are presented in Figures 5.19 and 5.20[82]. As shown in these figures, all listeners rated English foils' production of words as more native (mean range = 1.54 – 2.60), whereas learners' productions were judged as more foreign-accented (mean range = 3.60 – 6.80).

Averaged over the 241 subjects' productions of the six words, the judges' mean ratings were as follows: 4.69, 5.33, 5.12, 4.48, 4.40, and 4.56 for judges 1, 2, 3, 4, 5, and 6, respectively. Inter-rater reliability was calculated on the average accent ratings of the six words. A high degree of inter-listener agreement was reached; specifically, Cronbach's alpha was .94. As for ratings on each of the six words averaged over judges, means of 4.85, 4.84, 4.39, 4.43, 5.07, and 4.94 were obtained for the words *jam*, *reading*,

[81] When looking at learner groups in more detail, results for C1 and C2 will be only descriptive. That is, these two groups were not included in the statistical analyses conducted, due to the small sample of subjects comprising either group.

[82] For the sake of clarity, the accent ratings obtained for each language and gender subgroup have been pooled into a single rating averaged over judges and words. As will be discussed later, this approach – i.e. using a single mean rating – is not entirely satisfactory based on the significant results found in the repeated measures ANOVAs performed in order to reduce the amount of data.

red, *speak*, *this*, and *very*, respectively. In this case, Cronbach's alpha returned a value of .82, which indicated that the six words seemed to have been rated in a similar manner.

Further analyses were performed in an attempt to reduce data for the study of the possible effects of the research variables on the global FA ratings. Thus, the mean FA ratings given by every judge (averaged over words) were submitted to a (5) age of FL learning x (4) exposure x (6) listener repeated measures analysis of variance with repeated measures on listener. The simple effect for listener was significant (F (5, 217) = 38.961, $p < .001$), as well as the listener x age of FL learning (F (15, 599) = 5.707, $p < .001$) and listener x exposure (F (10, 434) = 3.806, $p < .001$) interactions, but not the three-way interaction (listener x age of FL learning x exposure) (F (10, 434) = 1.786, $p = .061$). As for the significant effect for listener, pairwise comparisons showed that judges 2's and 3's ratings were, on the whole, significantly more accented than those of the remaining judges (Bonferroni $p < .05$). On the other hand, judge 1 rated subjects' productions as significantly less accented than judges 2, 3, 4, and 5. Moreover, homogeneity tests (Levene's Tests of Equality of Error Variances) yielded a significant result for judge 5, which indicated that her ratings were not normally distributed (F (8, 221) = 2.126, $p < .05$).

Figure 5.17. Mean FA ratings for judges averaged over words.

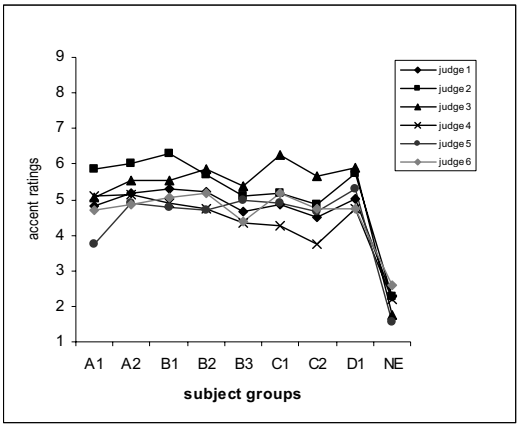


Figure 5.18. Mean FA ratings on words averaged over judges.

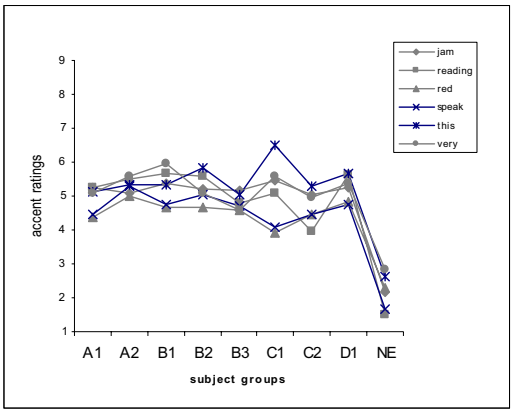


Figure 5.19. Mean global FA ratings for each language subgroup averaged across words and judges.

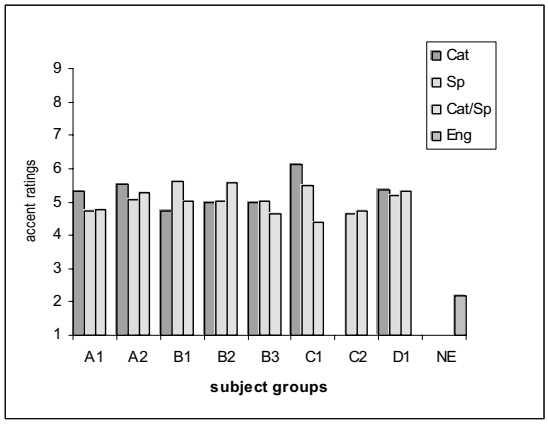
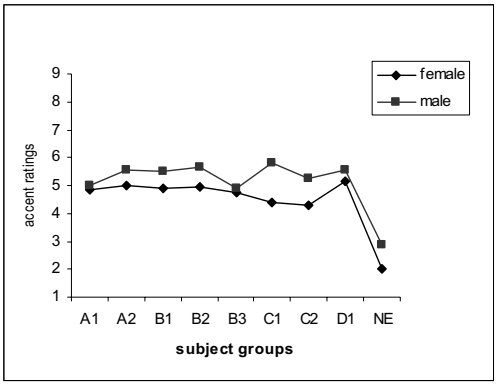


Figure 5.20. Mean global FA ratings for male and female subjects averaged across words and judges.



A further repeated measures ANOVA on the average ratings (across judges) on each of the six words with age of FL learning and exposure as between-subjects factors and word as a within-subjects factor revealed that ratings on words did differ one from another significantly (F (5, 217) = 7.621, $p < .001$). The word x exposure interaction reached significance, as well (F (10, 434) = 1.928, $p < .05$). No other two-way or three-way interaction yielded significant results (word x AOL [F (15, 599) = .678, $p = .807$], word x AOL x exposure [F (10, 434) = 1.061, $p = .392$]). Word pairwise comparisons showed that *jam*, *reading*, *this*, and *very* were judged as significantly more foreign-accented than *red* and *speak* (Bonferroni $p < .05$).

As in the previous analysis, tests of normality showed that in the words *jam* and *reading* ratings were not normally distributed (F (8, 221) = 2.453, $p < .05$; F (8, 221) = 2.733, $p < .05$).

Based on the significant results reported above, a single mean rating comprising judges' ratings and ratings averaged over words was not deemed satisfactory, statistically speaking. Furthermore, the violation of the assumption of homogeneity of variances in some instances, as well as the fact that the scale used was ordinal, made it advisable to opt for nonparametric procedures. Therefore, like the statistical analyses performed on subjects' discrimination scores, Kruskal-Wallis analyses were also computed in order to assess the effects of AOL, exposure, L1(s), and gender on the accent scores (averaged over judges or over words) obtained for the rating task. In the event of significant

differences found among groups, Mann-Whitney *U* tests were then carried out to determine where exactly the differences occurred. As this involved a large number of same subject-group comparisons, the alpha level was set at .001 in order for a result to be considered statistically significant at the .05 level (.001 x 6 words x 6 judges).

5.2.2.1.1. Effect of onset age of FL learning

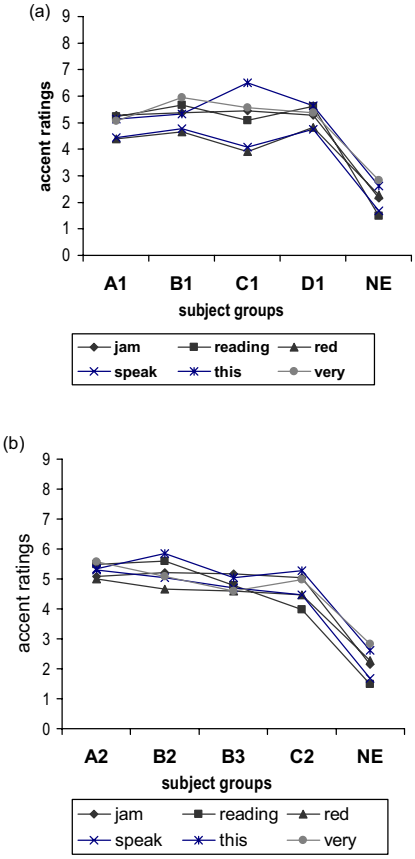
Separate Kruskal-Wallis analyses were performed, on the one hand, on the mean accent ratings averaged over judges, and, on the other hand, on the mean ratings averaged over words as dependent variables; and with age of FL learning as a factor. As stated above, in order to locate the resulting significant differences between groups, Mann-Whitney *U* tests were employed.

In all instances, Kruskal-Wallis analyses resulted in significant differences among groups (adjusted $p < .05$). More precisely, all learner groups' ratings differed significantly from those of English NSs, both when ratings referred to words and to judges (adjusted $p < .05$). As mentioned above, English foils received significantly lower scores than the various learner groups' ratings. In the latter groups, the accent scores obtained indicated that FL learners mostly produced the six words with a moderate amount of FA, according to all listeners (see Figures 5.21 and 5.22).

Among learner groups, though scores were often at the midpoint of the scale of FA, a few tendencies can be observed. For the most part, eight-year-old beginners with 200 hours of exposure received slightly less accented ratings on words and by all listeners than 11-year-old and adult starters matched for exposure (see Figures 5.21a and 5.22a). These differences in scores resulted in the following significant results and/or results that were close to being significant. First, the difference in ratings for A1 and B1 approached significance in the word *very* (*U* 264, *Z* –2.379, $p = .017$, adjusted $p > .05$) and in the ratings assigned by judges 1, 3, and 5 (*U* 280, *Z* –1.810, $p = .070$, adjusted $p > .05$; *U* 244.5, *Z* –2.396, $p = .017$, adjusted $p > .05$; and *U* 202.5, *Z* –3.085, $p = .002$, adjusted $p > .05$). And the between-comparison of A1–D1 was significant for judges 3 and 5 (*U* 216.5, *Z* –3.320, adjusted $p < .05$; *U* 162, *Z* –4.144, adjusted $p < .05$). In addition, judge 5's ratings for B1 vs. D1 were close to being significant (*U* 260, *Z* –1.975, $p = .048$, adjusted $p > .05$).

Age differences were not so clear-cut when learners had 416 hours of instruction in the FL. In the words *red*, *speak*, and *very*, 11-year-old beginners obtained lower accent scores (i.e. a weaker foreign accent) than 8-year-old beginners, whereas the reverse applied to the words *jam*, *reading*, and *this* (see Figure 5.21b). Similarly, judges 2, 4, and 5 rated B2's word productions as less foreign-accented as those of A2. And judges 1, 3, and 6 did so for A2 in relation to B2 (see Figure 5.22b). In no case were the differences observed large enough to be significant.

Figure 5.21. Mean FA ratings on words (averaged over judges). Factor: onset age of FL learning. Results for C1 and C2 are indicative only.



The effect of age of first exposure to English in different learner groups with 726 hours of instruction could not be tested, as there was only one group: B3. As mentioned above, when compared to NSs of English, Mann-Whitney *U* tests showed that B3's accent ratings were significantly more accented than those of the NE group (adjusted $p < .05$). Tables 5.16 and 5.17 below summarise all the comparisons carried out with AOL as a factor.

Figure 5.22. Judges' mean FA ratings (averaged over words). Factor: onset age of FL learning. Results for C1 and C2 are indicative only.

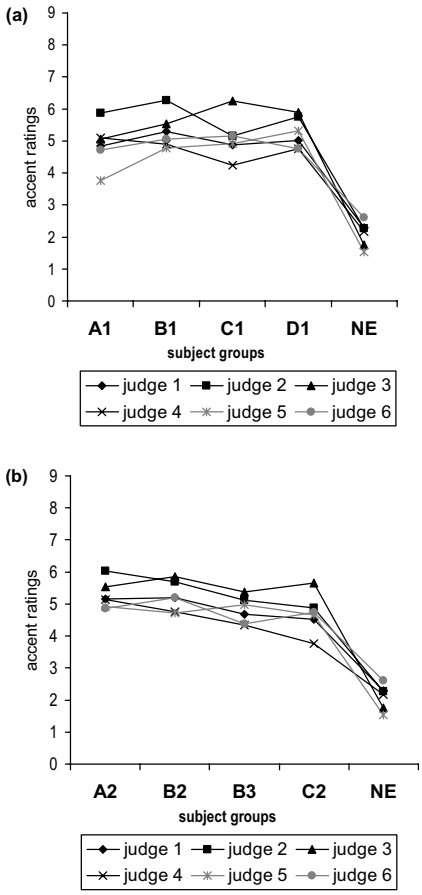


Table 5.16. Summary of comparisons carried out on the accent ratings on words (averaged over judges) with AOL as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as $p =$ .exact significance value. Each group's mean accent ratings for each word appear in parentheses.

(a)

Factor: ***onset age of FL learning***
D.V.: FA ratings on *jam*
A1 (5.27) – B1 (5.37) – D1 (5.27) – NE (2.15)*
A1 – B1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – D1 n.s.
B1 < NE*
D1 < NE*
A2 (5.09) – B2 (5.22) – NE (2.15)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

(b)

Factor: ***onset age of FL learning***
D.V.: FA ratings on *reading*
A1 (5.23) – B1 (5.67) – D1 (5.62) – NE (1.48)*
A1 – B1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – D1 n.s.
B1 < NE*
D1 < NE*
A2 (5.48) – B2 (5.59) – NE (1.48)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

(c)

Factor: ***onset age of FL learning***
D.V.: FA ratings on *red*
A1 (4.39) – B1 (4.66) – D1 (4.82) – NE (2.28)*
A1 – B1 n.s.
A1 > D1 ***p*** = .055
A1 < NE*
B1 – D1 n.s.
B1 < NE*
D1 < NE*
A2 (5.01) – B2 (4.65) – NE (2.15)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

(d)

Factor: ***onset age of FL learning***
D.V.: FA ratings on *speak*
A1 (4.44) – B1 (4.77) – D1 (4.75) – NE (1.68)*
A1 – B1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – D1 n.s.
B1 < NE*
D1 < NE*
A2 (5.29) – B2 (5.04) – NE (1.68)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

Table 5.16 (continued)

(e)

Factor: ***onset age of FL learning***
D.V.: FA ratings on *this*
A1 (5.13) – B1 (5.32) – D1 (5.65) – NE (2.61)*
A1 – B1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – D1 n.s.
B1 < NE*
D1 < NE*
A2 (5.34) – B2 (5.85) – NE (2.61)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

(f)

Factor: ***onset age of FL learning***
D.V.: FA ratings on *very*
A1 (5.07) – B1 (5.95) – D1 (5.37) – NE (2.84)*
A1 > B1 $p = .017$
A1 – D1 n.s.
A1 < NE*
B1 – D1 n.s.
B1 < NE*
D1 < NE*
A2 (5.57) – B2 (5.08) – NE (2.84)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p < .05$.

Table 5.17. Summary of comparisons carried out on the judges' accent ratings (averaged over words) with AOL as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as $p =$ .exact significance value. Each group's mean accent ratings by each judge appear in parentheses.

(a)

Factor: ***onset age of FL learning***
D.V.: Judge 1's FA ratings
A1 (4.83) – B1 (5.29) – D1 (5.02) – NE (2.29)*
A1 >B1 $\boldsymbol{p} = .070$
A1 – D1 n.s.
A1 < NE*
B1 – D1 n.s.
B1 < NE*
D1 < NE*
A2 (5.16) – B2 (5.20) – NE (2.29)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

(b)

Factor: ***onset age of FL learning***
D.V.: Judge 2's FA ratings
A1 (5.87) – B1 (6.28) – D1 (5.75) – NE (2.27)*
A1 – B1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – D1 n.s.
B1 < NE*
D1 < NE*
A2 (6.03) – B2 (5.70) – NE (2.27)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

Table 5.17 (continued)

(c)

Factor: ***onset age of FL learning***
D.V.: Judge 3's FA ratings
A1 (5.06) – B1 (5.54) – D1 (5.89) – NE (1.75)*
A1 > B1 ***p*** **=** .017
A1 > D1*
A1 < NE*
B1 – D1 n.s.
B1 < NE*
D1 < NE*
A2 (5.54) – B2 (5.85) – NE (1.75)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

(d)

Factor: ***onset age of FL learning***
D.V.: Judge 4's FA ratings
A1 (5.10) – B1 (4.89) – D1 (4.75) – NE (2.18)*
A1 – B1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – D1 n.s.
B1 < NE*
D1 < NE*
A2 (5.13) – B2 (4.75) – NE (2.18)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

(e)

Factor: ***onset age of FL learning***
D.V.: Judge 5's FA ratings
A1 (3.76) – B1 (4.77) – D1 (4.91) – NE (1.54)*
A1 > B1 ***p*** **=** .002
A1 > D1*
A1 < NE*
B1 > D1 ***p*** **=** .048
B1 < NE*
D1 < NE*
A2 (4.92) – B2 (4.71) – NE (1.54)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

(f)

Factor: ***onset age of FL learning***
D.V.: Judge 6's FA ratings
A1 (4.72) – B1 (5.05) – D1 (4.76) – NE (2.60)*
A1 – B1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – D1 n.s.
B1 < NE*
D1 < NE*
A2 (5.57) – B2 (5.08) – NE (2.60)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p < .05$.

5.2.2.1.2. Effect of exposure

As shown in Figures 5.23a and 5.24a, 8-year-old beginners with a greater amount of instruction (416 hours) generally received more foreign-accented ratings than learners with 200 hours of exposure. In fact, only A2's production of the word *jam* obtained less foreign-accented scores than A1, as an increase in exposure would lead one to expect. The observed differences in ratings between A1 and A2 were significant in the mean ratings by judge 5 (U 296.5, Z –3.813, p = .000, adjusted $p < .05$), and approached significance in the words *red* and *speak* (U 466, Z –2.382, p = .017, adjusted $p > .05$; U 412, Z –2.959, p = .003, adjusted $p > .05$) and in the mean ratings by judge 3 (U 431, Z –2.278, p = .023, adjusted $p > .05$).

In the case of 11-year-old starters, Kruskal-Wallis analyses revealed a significant effect of amount of instruction on the production of the six words as rated by judge 2 (χ^2 19.057, df 2, adjusted $p < .05$). In that case, B3 obtained significantly less accented scores than B1 (U 197, Z –4.159, p = .000, adjusted $p < .05$). The differences between B3 and B2, on the one hand, and between B2 and B1, on the other hand, approached significance (for B2–B3: U 338.5, Z –2.456, p = .014, adjusted $p > .05$; and for B1–B2: U 234.5, Z –2.078, p = .038, adjusted $p > .05$), all in favour of learners with a higher amount of instruction in English. Moreover, the ratings obtained for the words *reading* and *very* approached significance, as exposure increased (χ^2 7.651, df 2, p = .022 adjusted, $p > .05$; and χ^2 9.679, df 2, p = .008, adjusted $p > .05$), in addition to the ratings provided by judges 1 and 6 (χ^2 9.679, df 2, p = .012, adjusted $p > .05$; and χ^2 11.823, df 2, p = .003, adjusted $p > .05$).

Unlike 8-year-old beginners, on the whole, 11-year-old-starters' increase in experience led to less foreign-accented word productions. Besides, judges tended to rate subjects' words as less accented when they had more hours of formal instruction in English (Figures 5.23b and 5.24b). Exceptions to this pattern were found mainly in learners with 416 hours, who received higher ratings in the words *speak* and *this*, and in judges 3's and 6's ratings. This finding agrees with that of the comparison between A1's and A2's ratings, where A2 produced the six words with a higher degree of FA than A1.

At a descriptive level, the accent scores obtained for C1 and C2 showed that, generally, more exposure resulted in less accented ratings (Figures 5.23c and 5.24c).

Tables 5.18 and 5.19 present a summary of all the group comparisons carried out on the judges' FA ratings on words and with amount of instruction to English as a factor.

Figure 5.23. Mean FA ratings on words (averaged over judges). Factor: exposure to the FL. Results for C1 and C2 are indicative only.

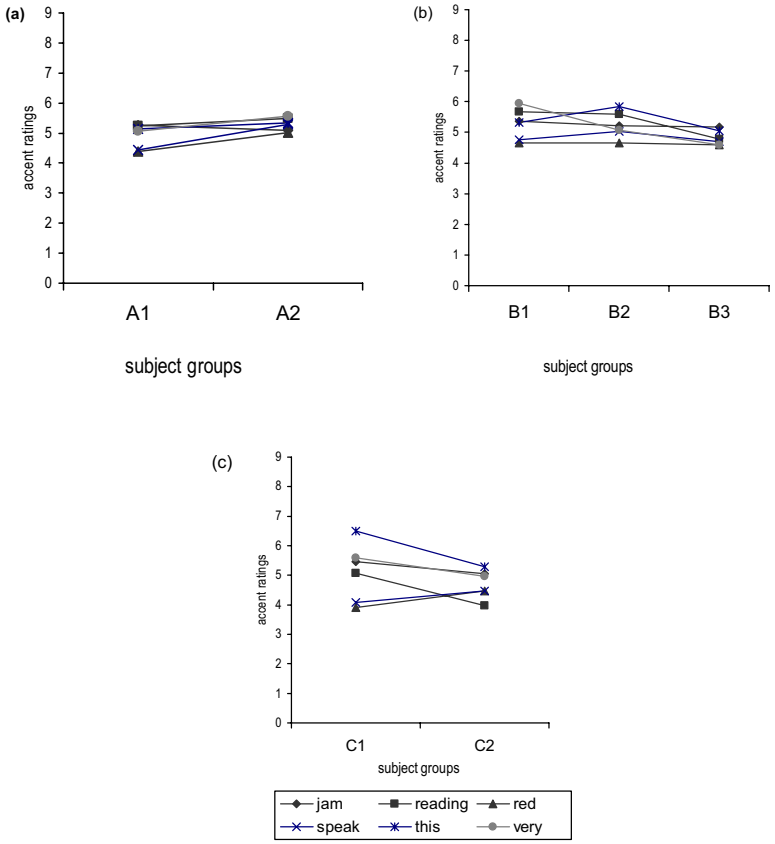


Figure 5.24. Judges' mean FA ratings (averaged over words). Factor: exposure to the FL. Results for C1 and C2 are indicative only.

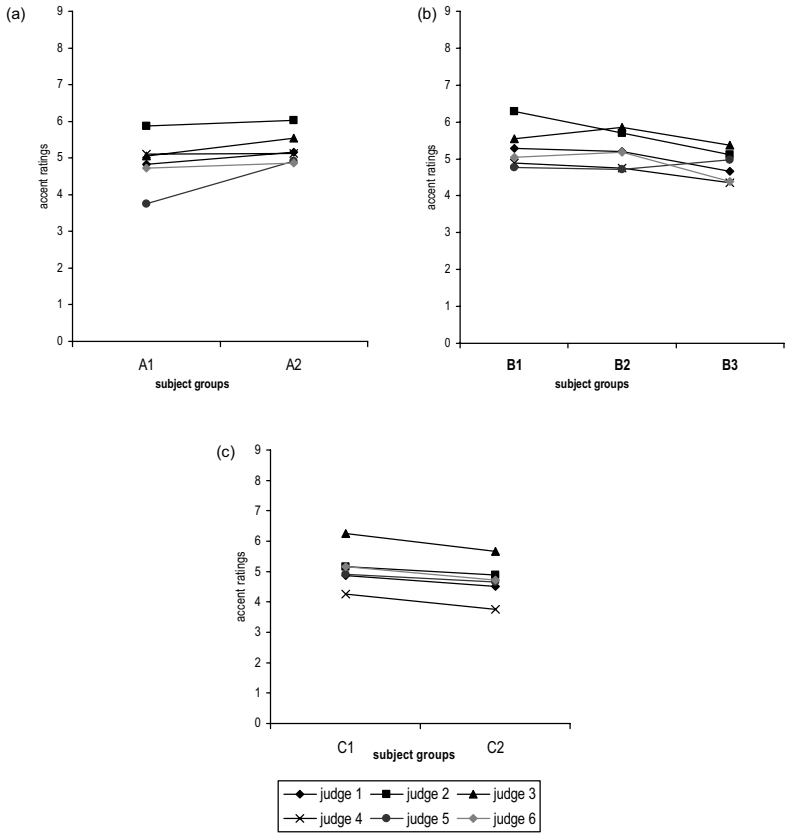


Table 5.18. Summary of comparisons carried out on the accent ratings on words (averaged over judges) with exposure to the FL as a factor. Nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = exact significance value. Each group's mean accent ratings for each word appear in parentheses.

(a)

Factor: ***exposure to English*** *D.V.:* FA ratings on *jam*
A1 (5.27) – A2 (5.09) n.s.
B1 (5.37) – B2 (5.22) – B3 (5.18) n.s.

(b)

Factor: ***exposure to English*** *D.V.:* FA ratings on *reading*
A1 (5.23) – A2 (5.48) n.s.
B1 (5.67) – B2 (5.59) – B3 (4.79) ***p*** = .022

(c)

Factor: ***exposure to English*** *D.V.:* FA ratings on *red*
A1 (4.39) > A2 (5.01) ***p*** = .017
B1 (4.66) – B2 (4.65) – B3 (4.60) n.s.

(d)

Factor: ***exposure to English*** *D.V.:* FA ratings on *speak*
A1 (4.44) > A2 (5.29) ***p*** = .003
B1 (4.77) – B2 (5.04) – B3 (4.70) n.s.

(e)

Factor: ***exposure to English*** *D.V.:* FA ratings on *this*
A1 (5.13) – A2 (5.34) n.s.
B1 (5.34) – B2 (5.85) – B3 (5.05) ***p*** = .083

(f)

Factor: ***exposure to English*** *D.V.:* FA ratings on *very*
A1 (5.07) – A2 (5.57) n.s.
B1 (5.95) – B2 (5.08) – B3 (4.59) ***p*** =.008

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p <$ .05.

Table 5.19. Summary of comparisons carried out on the judges' accent ratings (averaged over words) with exposure to the FL as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each group's mean accent ratings by each judge appear in parentheses.

(a)

Factor: ***exposure to English***
D.V.: Judge 1's FA ratings
A1 (4.83) – A2 (5.16) n.s.
B1 (5.29) – B2 (5.20) – B3 (4.67) ***p*** **=** .012

(b)

Factor: ***exposure to English***
D.V.: Judge 2's FA ratings
A1 (5.87) – A2 (6.03) n.s.
B1 (6.28) – B2 (5.70) – B3 (5.11)* B1 < B2 ***p*** **=** .038 B1 < B3* B2 < B3 ***p*** **=** .014

(c)

Factor: ***exposure to English***
D.V.: Judge 3's FA ratings
A1 (5.06) > A2 (5.54) ***p*** **=** .023
B1 (5.54) – B2 (5.85) – B3 (5.38) n.s.

(d)

Factor: ***exposure to English***
D.V.: Judge 4's FA ratings
A1 (5.10) – A2 (5.13) n.s.
B1 (4.89) – B2 (4.75) – B3 (4.35) n.s.

(e)

Factor: ***exposure to English***
D.V.: Judge 5's FA ratings
A1 (3.76) > A2 (4.92)*
B1 (4.77) – B2 (4.71) – B3 (4.97) n.s.

(f)

Factor: ***exposure to English***
D.V.: Judge 6's FA ratings
A1 (4.72) – A2 (4.86) n.s.
B1 (5.05) – B2 (5.19) – B3 (4.39) ***p*** **=** .003

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p < .05$.

5.2.2.1.3. Effect of dominant L1(s)

Kruskal-Wallis analyses with L1(s) as a factor yielded significant differences among groups in their productions of all the six words and for all the six judges (adjusted $p < .05$) (see Figures 5.25 and 5.26). Mann-Whitney *U* tests indicated that in all cases the NE group received significantly lower accent ratings (i.e. more native-like) than the three language subgroups (adjusted $p < .05$). Among NNSs, no language subgroup received either significantly less or more foreign-accented ratings than the remaining two language groups. Only on one occasion did a between-group comparison approach significance – namely, Catalan dominant speakers' higher foreign-accented scores than Catalan/Spanish balanced bilinguals in the production of the word *very* (M = 5.62 vs. 5.09) (U 1883, Z –1.941, p = .052, adjusted $p > .05$). Tables 5.20 and 5.21 display all the comparisons conducted with the factor of dominant L1(s).

Figure 5.25. Mean FA ratings on words (averaged over judges). Factor: dominant L1(s).

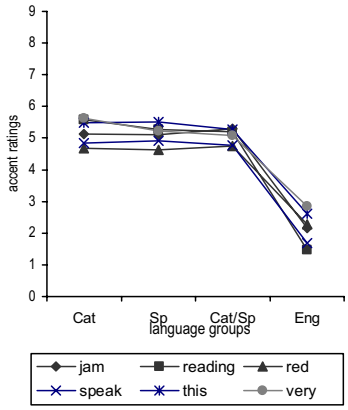


Figure 5.26. Judges' mean FA ratings (averaged over words). Factor: dominant L1(s).

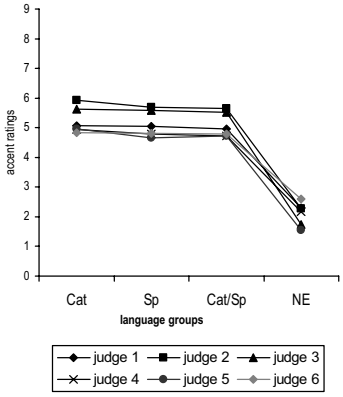


Table 5.20. Summary of comparisons carried out on the accent ratings on words (averaged over judges) with dominant L1(s) as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each language group's mean accent ratings for each word appear in parentheses.

(a)

Factor: ***dominant L1(s)***
D.V.: FA ratings on *jam*
Cat (5.12) – Sp (5.10) – C/S (5.30) – NE (2.15)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(b)

Factor: ***dominant L1(s)***
D.V.: FA ratings on *reading*
Cat (5.58) – Sp (5.28) – C/S (5.21) – NE (1.48)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(c)

Factor: ***dominant L1(s)***
D.V.: FA ratings on *red*
Cat (4.67) – Sp (4.64) – C/S (4.75) – NE (2.28)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(d)

Factor: ***dominant L1(s)***
D.V.: FA ratings on *speak*
Cat (4.85) – Sp (4.92) – C/S (4.78) – NE (1.68)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(e)

Factor: ***dominant L1(s)***
D.V.: FA ratings on *this*
Cat (5.48) – Sp (5.50) – C/S (5.28) – NE (2.61)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(f)

Factor: ***dominant L1(s)***
D.V.: FA ratings on *very*
Cat (5.62) – Sp (5.22) – C/S (5.09) – NE (2.84)*
Cat – Sp n.s.
Cat < C/S ***p*** = .052
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p < .05$.

Table 5.21. Summary of comparisons carried out on the judges' accent ratings (averaged over words) with dominant L1(s) as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). Each language group's mean accent ratings by each judge appear in parentheses.

(a)

Factor: ***dominant L1(s)***
D.V.: Judge 1's FA ratings
Cat (5.06) – Sp (5.04) – C/S (4.96) – NE (2.29)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(b)

Factor ***dominant L1(s)***
D.V.: Judge 2's FA ratings
Cat (5.93) – Sp (5.69) – C/S (5.64) – NE (2.27)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(c)

Factor: ***dominant L1(s)***
D.V.: Judge 3's FA ratings
Cat (5.63) – Sp (5.59) – C/S (5.51) – NE (1.75)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(d)

Factor: ***dominant L1(s)***
D.V.: Judge 4's FA ratings
Cat (4.94) – Sp (4.79) – C/S (4.72) – NE (2.18)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(e)

Factor: ***dominant L1(s)***
D.V.: Judge 5's FA ratings
Cat (4.96) – Sp (4.66) – C/S (4.72) – NE (1.54)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

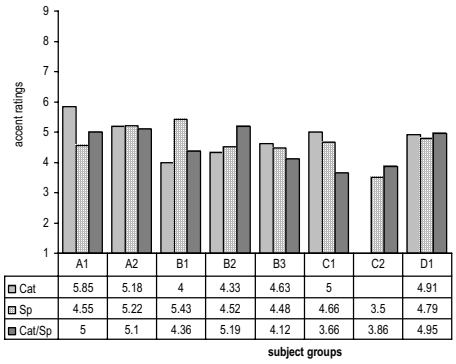
(f)

Factor: ***dominant L1(s)***
D.V.: Judge 6's FA ratings
Cat (4.83) – Sp (4.82) – C/S (4.79) – NE (2.60)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$.

When each learner group was considered separately, there was no consistent pattern as to what language subgroup received higher or lower or the same scores across words and judges. For instance, in the group of earlier starters with 200 hours of exposure (A1), Catalan dominant speakers obtained noticeably higher accent ratings than both Spanish dominant speakers and balanced bilinguals in the word *very*, as well as for judge 4. These differences only approached significance (*U* 11, *Z* –2.928, *p* = .003, adjusted *p* > .05; and *U* 14, *Z* –2.539, *p* = .011, adjusted *p* > .05, respectively). By contrast, among 11-year-old starters with 200 hours of exposure (B1), judge 4 rated Spanish dominant speakers' productions as noticeably more foreign-accented than balanced bilinguals (*U* 32, *Z* –2.098, *p* = .036, adjusted *p* > .05). Figure 5.27[83] illustrates this finding of no consistent pattern among learner groups. Besides, it is worth mentioning the fact that judges 5's and 6's somewhat higher degree of familiarity with the Spanish language did not influence their ratings in a noticeable way (e.g. by considering Spanish dominant speakers' productions as consistently either more foreign-accented or less foreign-accented than the other language subgroups).

Figure 5.27. Judge 4's mean FA ratings on words. Factor: dominant L1(s). Results for C1 and C2 are indicative only.



[83] Due to the large number of dependent variables and subgroups within each learner group, only a small, representative, number of figures are presented in this section (5.2.2.1.3) and the following section (5.2.2.1.4). On the same grounds, summary tables of the comparisons carried out within each learner group are not presented (for they did not yield any significant results, either).

5.2.2.1.4. Effect of gender

Overall, the FA ratings obtained for female subjects were lower than those of male subjects[84], as shown in Figures 5.28 and 5.29. This resulted in significant differences in the ratings for *red* and *speak* (U 4026, Z –3.298, adjusted $p < .05$; and U 3844, Z – 2.770, adjusted $p < .05$), in addition to the ratings assigned by judges 1, 3, and 6 to female and male speakers (U 3230, Z –4.522, adjusted $p < .05$; U 3128, Z –4.763, adjusted $p < .05$; and U 3696, Z –3.405, adjusted $p < .05$, respectively). In the remaining cases where there were no significant differences, the direction of the effect was the same as for the other words and judges, but it approached, rather than reaching, significance (adjusted $p > .05$) (see Tables 5.22 and 5.23 for a summary of all the comparisons carried out).

Figure 5.28. Mean FA ratings on words (averaged over judges). Factor: gender.

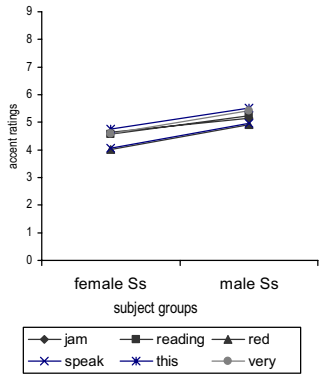
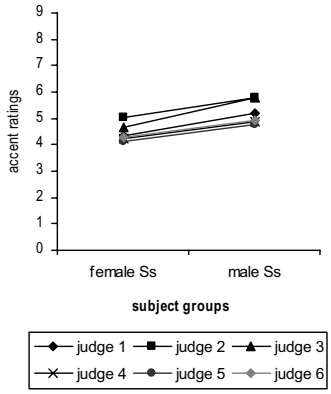


Figure 5.29. Judges' mean FA ratings (averaged over words). Factor: gender.



[84] English-speaking subjects were not included in this analysis, for the performance of male learners vs. female learners in an instructional setting was, in this case, of greater interest/importance. In spite of this, preliminary analysis of English NSs' production of English words showed that female speakers also received less foreign-accented ratings than male speakers (see also Figure 5.20 in section 5.2.2.1 above).

Table 5.22. Summary of comparisons carried out on the accent ratings on words (averaged over judges) with gender as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each gender group's mean accent ratings for each word appear in parentheses.

(a)

Factor: ***gender***
D.V.: FA ratings on *jam*
Female (4.64) – Male (5.14) n.s.

(b)

Factor: ***gender***
D.V.: FA ratings on *reading*
Female (4.57) – Male (5.25) n.s.

(c)

Factor: ***gender***
D.V.: FA ratings on *red*
Female (4.02) > Male (4.92)*

(d)

Factor: ***gender***
D.V.: FA ratings on *speak*
Female (4.07) > Male (4.96)*

(e)

Factor: ***gender***
D.V.: FA ratings on *this*
Female (4.76) > Male (5.51) ***p*** = .006

(f)

Factor: ***gender***
D.V.: FA ratings on *very*
Female (4.60) > Male (5.43) ***p*** = .013

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 and the resulting adjusted $p < .05$.

Table 5.23. Summary of comparisons carried out on the judges' accent ratings (averaged over words) with gender as a factor. Significant comparisons are marked with * ($p < .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each gender group's mean accent ratings by each judge appear in parentheses.

(a)

Factor: ***gender***
D.V.: Judge 1's FA ratings
Female (4.33) > Male (5.21)*

(b)

Factor ***gender***
D.V.: Judge 2's FA ratings
Female (5.02) > Male (5.79) ***p*** = .013

(c)

Factor: ***gender***
D.V.: Judge 3's FA ratings
Female (4.68) > Male (5.76)*

(d)

Factor: ***gender***
D.V.: Judge 4's FA ratings
Female (4.23) > Male (4.85) ***p*** = .016

(e)

Factor: ***gender***
D.V.: Judge 5's FA ratings
Female (4.14) > Male (4.76) ***p*** = .096

(f)

Factor: ***gender***
D.V.: Judge 6's FA ratings
Female (4.29) > Male (4.95)*

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p < .05$.

When the file was split by subject group, the following was observed. First, male and female subjects in A1, B3, and D1 obtained very similar ratings on every word (for an example, see Figure 5.30). As for the remaining learner groups, male and female production of English words was more distinct, since female subjects normally obtained less foreign-accented scores. Further, the differences in FA ratings between male and female speakers were larger in 11-year-old beginners with 200 hours and 416 hours of instruction (B1 and B2) than in the other subject groups. None of the differences in ratings between male and female speakers was significant, according to Mann-Whitney *U* tests (adjusted $p > .05$), though, in the case of B1 and B2, some differences approached significance (see Figure 5.31 as an example).

Figure 5.30. Example of similarity in ratings between male and female speakers in A1, B3, and D1. Results for C1 and C2 are indicative only. Mean FA ratings on *very*.

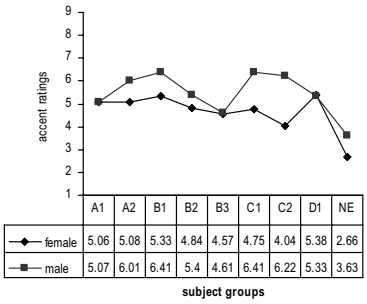
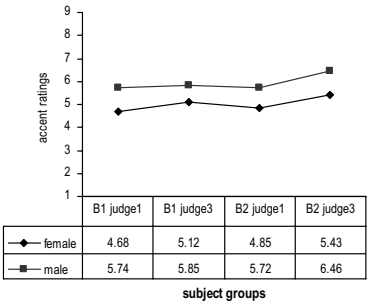


Figure 5.31. Example of differences between ratings for female and male speakers in B1 and B2 approaching significance: Judges 1's and 3's mean FA ratings on words.



5.2.2.2. FA ratings on segments

Frequencies of the scale points used by all judges to rate each segment indicated that judges generally followed the instructions given at the beginning of each block. Only on rating /i/ did judge 6 use six points on the scale (1–6). A few times the points used on the scale ranged from 1 to 7 (judges 3's and 6's ratings on /ɪ/ and /æ/). But, on the whole, judges used the 9 points on the scale in each rating session.

Mean FA ratings averaged over judges were obtained for the vowel segments, on the one hand, and for the consonant segments, on the other hand[85] (see Figures 5.32 and 5.33). In addition, judges' mean FA ratings averaged over vowel sounds as well as over consonant segments were calculated (Figures 5.34 and 5.35). As was the case of degree of global FA on words, English foils' segment productions were rated as more native-like

[85] Like the perceptual task, consonants and vowels were studied separately for two reasons. First, in L2 acquisition research these two types of sounds are often examined in different tasks. And, second, a Wilcoxon Signed Ranks test performed on the FA ratings for vowel and consonant segments (M = 3.51 and 4.59, respectively) showed that both sound types received significantly different accent ratings ($p < .001$). In this case, subjects' production of consonant segments presented a significant higher degree of FA than vowels.

(mean range = 1.19 – 2.52) than those of FL learners. NNSs' productions of vowel and consonant sounds ranged from a mean low of 2.56 to a mean high of 6.50.

Reliability coefficients were calculated on the following variables: mean FA ratings on vowels averaged across judges and mean FA ratings on consonants averaged across judges. The coefficients obtained were .47 and .37 for vowels and consonants, respectively. Coefficients were also computed on every judge's mean FA ratings averaged over vowel segments and each judge's FA ratings averaged over consonant segments, returning values of .84 and .92, respectively. Thus, while inter-rater reliability was high (.84 and .92), both vowels and consonants were rated on a varying basis (.47 and .37).

Mean FA ratings on vowels (averaged over judges) were submitted to a (5) age of FL learning x (4) exposure x (3) vowel segments repeated measures ANOVA with repeated measures on vowel segments. The simple effect for vowel segment was significant (F (2, 223)= 15.606, $p < .001$), but no other simple effects or two-way or three-way interactions yielded significant results. Pairwise comparisons with Bonferroni adjustments revealed that /i/ received significantly lower accent ratings than /ɪ/ and /æ/ ($p < .05$). Homogeneity tests also showed that ratings on the three vowel sounds were not normally distributed ($p < .05$). Another repeated measures ANOVA with the same between-subjects factors as above, but with mean FA ratings on consonant segments as within-subjects factors also yielded a significant effect for consonant segments (F (2, 226)= 72.831, $p < .001$), as well as significant two-way interactions: consonant segments x AOL (F (6, 452)= 3.929, $p < .001$), and consonant segments x exposure (F (4, 452)= 5.736, $p < .001$); and a significant three-way interaction: consonant segments x AOL x exposure (F (4, 452)= 1.890, $p < .001$). Pairwise comparisons between the three consonant segments under study indicated that the ratings obtained for each consonant segment differed significantly one from another (Bonferroni $p < .05$). From significantly less to more foreign-accented, the consonant sounds were: /s/, /v/, and /d/. Homogeneity tests were significant, as well ($p < .05$).

Based on the results from the reliability coefficients and repeated measures ANOVAs, no single average rating could be calculated among the consonant segments and vowel sounds.

Figure 5.32. Mean FA ratings on vowel segments (averaged over judges).

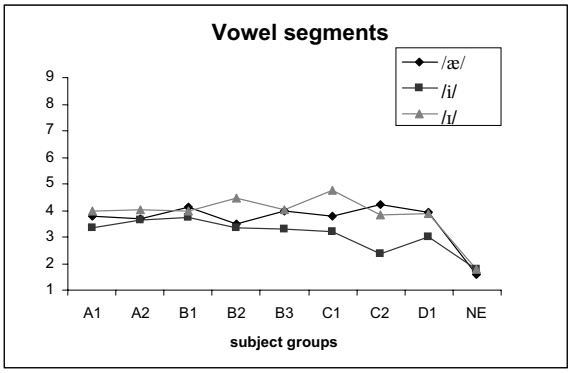
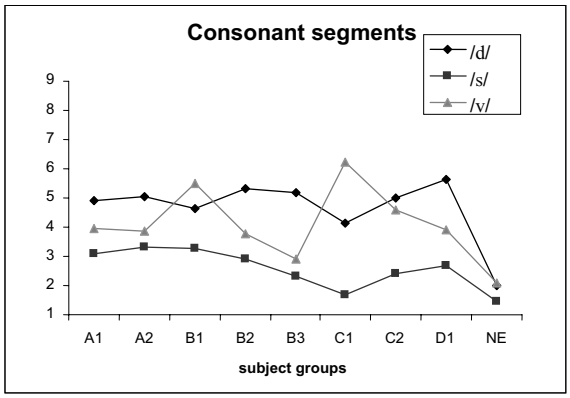


Figure 5.33. Mean FA ratings on consonant segments (averaged over judges).



Although inter-rater reliability coefficients were high, repeated measures analyses of variance were carried out on the average accent scores (averaged over consonant and vowel segments, respectively) for the six judges as repeated measures, and with AOL and exposure as factors. The simple effect for listener was significant in both analyses (F (5, 220)= 150.614, p < .001 for judges' ratings averaged over vowel segments; and F (5, 223)= 50.808, p < .001 for judges' ratings averaged over consonant segments). Homogeneity tests were significant in the average vowel ratings of judges 1, 2, 5, and 6.

For judges' average consonant ratings, homogeneity tests were nonsignificant, meaning that ratings were normally distributed.

Taking into account all the results reported above, nonparametric tests were preferred to parametric procedures to study the possible effects of AOL, exposure, dominant L1(s), and gender on the FA ratings obtained for segment production. Due to the large number of tests comparing the same subject groups, the significance level was set at .001 to maintain an experiment-wise error of .05 (.001 x 6 judges x 3 segments x 2 types of segment – vowel and consonant).

Figure 5.34. Judges' mean FA ratings on vowel segments (averaged over vowels).

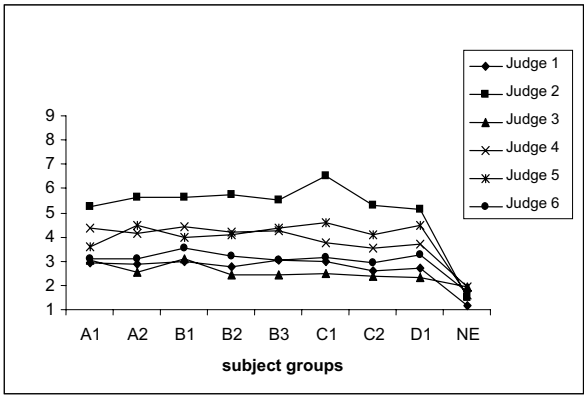
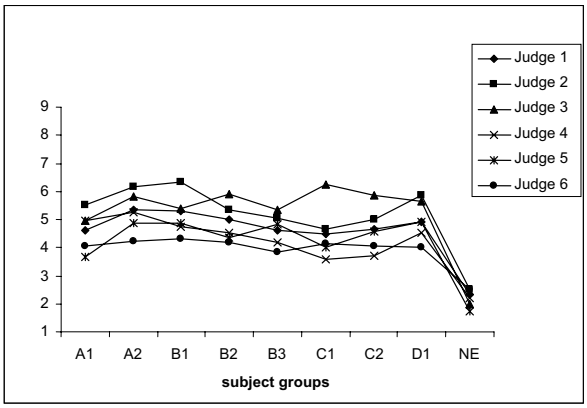


Figure 5.35. Judges' mean FA ratings on consonant segments (averaged over consonants).



5.2.2.2.1. Effect of age of FL learning

5.2.2.2.1.2. Production of vowel segments /i, ɪ, æ/

Mean FA ratings on the three vowel segments (averaged over judges) were submitted to Kruskal-Wallis analyses with age of FL learning as a factor. These analyses showed that significant differences existed in the accent ratings among groups (adjusted $p < .05$). Mann-Whitney *U* tests revealed that NE foils received lower ratings – to be exact, native-like – on the production of all /i/, /ɪ/ and /æ/ that differed significantly from those of learner groups (adjusted $p < .05$) (see Figure 5.36).

Among learner groups, a great deal of variability was observed as to which age group received lower or higher FA ratings, and depending on which vowel segment was examined. For instance, 8-year-old starters with 200 hours of instruction received lower scores for /i/ and /æ/ than 11-year-old beginners matched for amount of exposure, while for /ɪ/ A1 received higher scores than B1 (Figure 5.36a). On the contrary, with 416 hours of instruction, 8-year-old beginners obtained higher accent ratings for /i/ and /æ/, and lower ratings for /ɪ/ than 11-year-old beginners (Figure 5.36b). Likewise, and at a descriptive level, 14-year-old starters differed in the ratings obtained for /i/, /ɪ/ and /æ/ as a result of the amount of exposure and vowel sound being rated. Finally, adult learners produced vowel segments as less foreign-accented than younger starters.

Of all of the above observed differences (and variability in ratings), no group comparison reached significance, though two between-group comparisons approached significance – and always in the production of /i/: A1–D1 (U 300.5, Z –2.211, $p = .027$, adjusted $p > .05$) and B1–D1 (U 227.5, Z –2.697, $p = .007$, adjusted $p > .05$). In this case, differences were in favour of older starters (see Table 5.24).

Figure 5.36. Mean FA ratings on /i/, /ɪ/ and /æ/ averaged over judges. Factor: Onset age of FL learning. Results for C1 and C2 are indicative only.

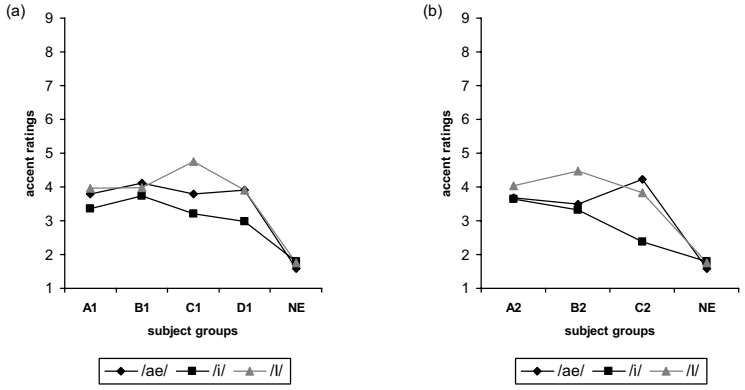


Table 5.24. Summary of comparisons carried out on the accent ratings on vowel segments (averaged over judges) with AOL as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each group's mean accent ratings for each vowel segment appear in parentheses.

(a)

Factor: ***onset age of FL learning***   *D.V.:* FA ratings on /i/
A1 (3.36) – B1 (3.73) – D1 (2.98) – NE (1.79)*   A1 – B1 n.s.   A1 < D1 ***p*** = .027   A1 < NE*   B1 < D1 ***p*** = .007   B1 < NE*   D1 < NE*
A2 (3.65) – B2 (3.32) – NE (1.79)*   A2 – B2 n.s.   A2 < NE*   B2 < NE*

(b)

Factor: ***onset age of FL learning***   *D.V.:* FA ratings on /ɪ/
A1 (3.96) – B1 (3.99) – D1 (3.90) – NE (1.76)*   A1 – B1 n.s.   A1 – D1 n.s.   A1 < NE*   B1 – D1 n.s.   B1 < NE*   D1 < NE*
A2 (4.03) – B2 (4.48) – NE (1.76)*   A2 – B2 n.s.   A2 < NE*   B2 < NE*

Table 5.24. (continued)

(c)

Factor: ***onset age of FL learning***   *D.V.:* FA ratings on /æ/
A1 (3.80) – B1 (4.12) – D1 (3.91) – NE (1.58)*   A1 – B1 n.s.   A1 > D1 n.s.   A1 < NE*   B1 – D1 n.s.   B1 < NE*   D1 < NE*
A2 (3.67) – B2 (3.50) – NE (1.58)*   A2 – B2 n.s.   A2 < NE*   B2 < NE*

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p < .05$.

Kruskal-Wallis analyses on each judge's mean FA ratings averaged over vowel segments as the dependent variable and age of FL learning as a factor also yielded significant differences in the ratings among groups (adjusted $p < .05$), with the sole exception of judge 3 when comparing A2, B2, and NE (i.e. learners with different starting ages but with the same amount of formal exposure – 416 hours) (χ^2 7.134, *df* 2, $p = .028$, adjusted $p > .05$).

With 200 hours of instruction, learner groups received significantly higher accent ratings than NSs of English by nearly all judges, but for judge 3's ratings for D1 (*U* 276, *Z* –1.911, $p = .056$). With 416 hours of instruction, all learner groups received significantly more-accented scores in comparison to NE subjects by all judges (adjusted $p < .05$), with the above-mentioned exception of judge 3 (see Figure 5.37).

When learners had 200 hours of exposure, all listeners rated young child beginners' (A1) vowel productions as less foreign-accented than older child starters' (B1). As for the remaining age groups, judges were not consistent in their ratings in relation to younger learners. As seen in Figure 5.37, adult learners obtained lower FA ratings than A1 and B1 according to judges 1, 2, 3, and 4, whereas judges 5 and 6

assigned D1 higher ratings in comparison to A1. In spite of this, a rating pattern emerged from listeners – i.e. listeners were fairly consistent in the degree of FA to which they differed from each other (as shown in Figure 5.37b). For example, judges 1 and 3 assigned ratings within the range 2.1 – 2.9, while judge 2 used the point 5 on the scale. That is to say, despite the fact that judges did not always agree on the ratings assigned to older learners (vs. young learners), they were internally consistent with respect to their use of ratings which differed by one or two points on the FA scale.

As far as learner groups are concerned, no significant differences were found among the various learner groups in any of the listeners' ratings (see Table 5.25). Only the following comparisons approached significance: A1–D1 for judges 3, 4, and 5 (U 279.5, Z –2.378, p = .017, adjusted p > .05; U 313.5, Z –1.852, p = .064, adjusted p > .05; and U 297, Z –2.104, p = .035, adjusted p > .05); and B1–D1 for judges 2, 3, and 4 (U 271, Z –1.798, p = .072, adjusted p > .05; U 210.5, Z –2.830, p = .005, adjusted p > .05; and U 276.5, Z –1.700, p = .089, adjusted p > .05).

Table 5.25. Summary of comparisons carried out on the judges' accent ratings (averaged over vowel segments) with AOL as a factor. Significant comparisons are marked with * (p < .05), while nonsignificant group comparisons are displayed as n.s. (p > .05). P values of marginally significant results (.05 < p < .10) are stated as p = .exact significance value. Each group's mean accent ratings by each judge appear in parentheses.

(a)

Factor: ***onset age of FL learning***
D.V.: Judge 1's FA ratings on vowels
A1 (2.91) – B1 (3.00) – D1 (2.07) – NE (1.19)*
A1 – B1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – D1 n.s.
B1 < NE*
D1 < NE*
A2 (2.88) – B2 (2.79) – NE (1.19)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

(b)

Factor: ***onset age of FL learning***
D.V.: Judge 2's FA ratings on vowels
A1 (5.23) – B1 (5.63) – D1 (5.13) – NE (1.50)*
A1 – B1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 < D1 ***p*** = .072
B1 < NE*
D1 < NE*
A2 (5.65) – B2 (5.75) – NE (1.50)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

Table 5.25 (continued)

(c)

Factor: ***onset age of FL learning***
D.V.: Judge 3's FA ratings on vowels

A1 (3.02) – B1 (3.11) – D1 (2.34) – NE (1.93)*
A1 – B1 n.s.
A1 < D1 ***p*** = .017
A1 < NE*
B1 < D1 ***p*** = .005
B1 < NE*
D1 < NE ***p*** = .056

A2 (2.52) – B2 (2.46) – NE (1.93) ***p*** = .028

(d)

Factor: ***onset age of FL learning***
D.V.: Judge 4's FA ratings on vowels

A1 (4.34) – B1 (4.41) – D1 (3.68) – NE (1.93)*
A1 – B1 n.s.
A1 < D1 ***p*** = .064
A1 < NE*
B1 < D1 ***p*** = .089
B1 < NE*
D1 < NE*

A2 (4.13) – B2 (4.22) – NE (1.93)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

(e)

Factor: ***onset age of FL learning***
D.V.: Judge 5's FA ratings on vowels

A1 (3.61) – B1 (3.98) – D1 (4.48) – NE (1.60)*
A1 – B1 n.s.
A1 > D1 ***p*** = .035
A1 < NE*
B1 – D1 n.s.
B1 < NE*
D1 < NE*

A2 (4.47) – B2 (4.08) – NE (1.60)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

(f)

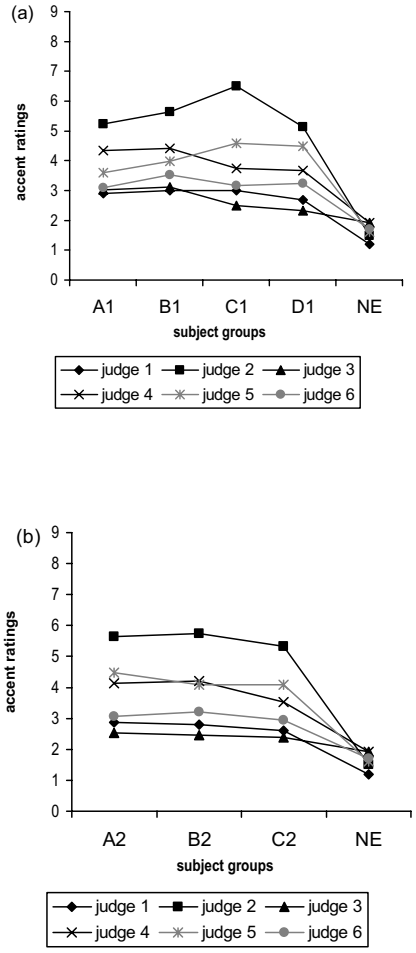
Factor: ***onset age of FL learning***
D.V.: Judge 6's FA ratings on vowels

A1 (3.10) – B1 (3.52) – D1 (3.24) – NE (1.71)*
A1 – B1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – D1 n.s.
B1 < NE*
D1 < NE*

A2 (3.07) – B2 (3.20) – NE (1.71)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p < .05$.

Figure 5.37. Judges' mean FA ratings on /i/, /ɪ/ and /æ/ (averaged over vowel segments). Factor: Onset age of FL learning. Results for C1 and C2 are indicative only.

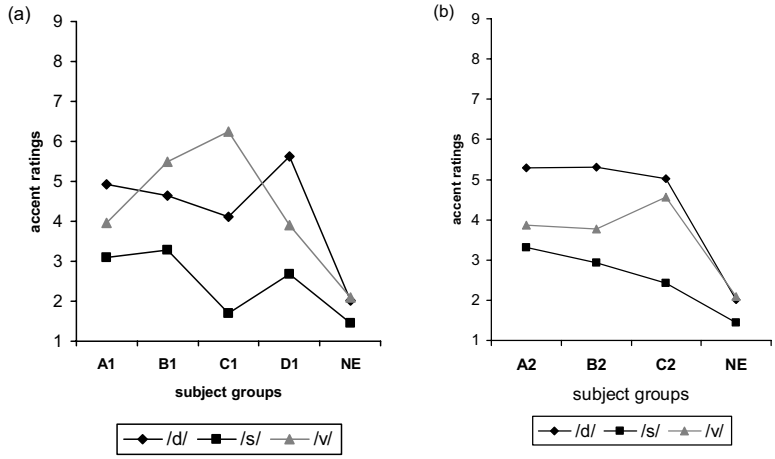


5.2.2.2.1.2. Production of consonant segments /d, s, v/

The mean FA ratings obtained for consonants – /d/, /s/, and /v/ – averaged over judges resulted in significant differences among all groups, according to Kruskal-Wallis analyses (adjusted $p < .05$). Mann-Whitney *U* tests showed that learner groups' ratings differed significantly from those of the NE group.

Among learner groups, the ratings obtained for the several age groups varied considerably depending on the consonant being examined (see Figure 5.38 for the mean accent ratings obtained). Thus, there was no well-defined advantage of one age group over the other. Nevertheless, in the cases where between-group comparisons approached significance, they were in favour of older learners (group D). Therefore, the following pairwise comparisons were close to being significant: A1–B1 in /v/ (*U* 267, *Z* –2.331, $p = .020$, adjusted $p > .05$); A1–D1 in /d/ (*U* 340, *Z* –1.793, $p = .073$, adjusted $p > .05$); and B1–D1 in /d/ and /v/ (*U* 251, *Z* –2.306, $p = .021$, adjusted $p > .05$; and *U* 242.5, *Z* –2.268, $p = .023$, adjusted $p > .05$, respectively). There were no significant differences among learner groups with 416 hours of exposure (see Table 5.26).

Figure 5.38. Mean FA ratings on /d/, /s/, and /v/ (averaged over judges). Factor: Onset age of FL learning. Results for C1 and C2 are indicative only.



The six judges' accent ratings averaged over consonant segments were also submitted to Kruskal-Wallis analyses with age of FL learning as a factor. Learner groups were found to produce consonant segments as significantly more foreign-accented than English foils (adjusted $p < .05$).

Age differences in ratings were not always consistent among judges when learners were at T1 (Figure 5.39a). On the other hand, judges tended to rate older learners' consonants as less foreign-accented than those of younger learners when they were at T2 (Figure 5.39b). According to Mann-Whitney *U* tests, judges 1's, 2's, and 5's accent ratings obtained for A1 and B1 were close to being significant (*U* 282.5, *Z* –2.094, $p = .036$, adjusted $p > .05$; *U* 270, *Z* –2.296, $p = .022$, adjusted $p > .05$; *U* 238, *Z* –2.789, $p = .005$, $p < .05$), as well as judges 3's and 5's ratings for A1 and D1 (*U* 308, *Z* –2.266, $p = .023$, adjusted $p > 05$; *U* 265.5, *Z* –2.872, $p = .004$, adjusted $p > .05$). And with learners with 416 hours of instruction the differences in accent scores that approached significance were located between A2 and B2 for judges 2 and 4 (*U* 378.5, *Z* –2.638, $p = .008$, adjusted $p > .05$; *U* 415.5, *Z* –2.204; $p = .028$, $p < .05$) (see Table 5.27).

Figure 5.39. Judges' mean FA ratings on /d/, /s/ and /v/ (averaged over consonant segments). Factor: Onset age of FL learning. Results for C1 and C2 are indicative only.

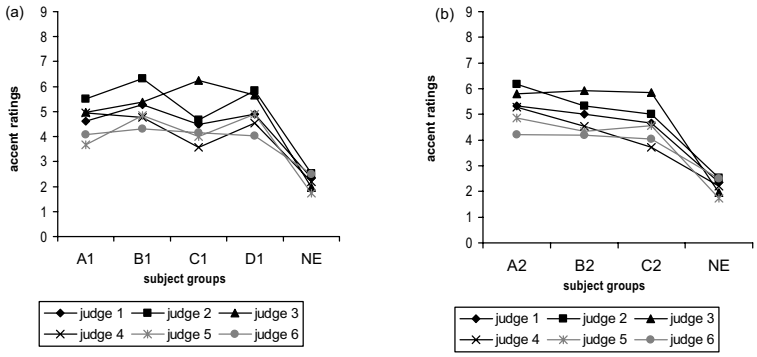


Table 5.26. Summary of comparisons carried out on the accent ratings on consonant segments (averaged over judges) with AOL as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each group's mean accent ratings for each consonant segment appear in parentheses.

(a)

Factor: ***onset age of FL learning***
D.V.: FA ratings on /d/

A1 (4.92) – B1 (4.64) – D1 (5.62) – NE (2.02)*
A1 – B1 n.s.
A1 > D1 ***p*** = .073
A1 < NE*
B1 > D1 ***p*** = .021
B1 < NE*
D1 < NE*

A2 (5.03) – B2 (5.31) – NE (2.02)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

(b)

Factor: ***onset age of FL learning***
D.V.: FA ratings on /s/

A1 (3.10) – B1 (3.29) – D1 (2.68) – NE (1.45)*
A1 – B1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – D1 n.s.
B1 < NE*
D1 < NE*

A2 (3.31) – B2 (2.93) – NE (1.45)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

(c)

Factor: ***onset age of FL learning***
D.V.: FA ratings on /v/

A1 (3.96) – B1 (5.50) – D1 (3.90) – NE (2.10)*
A1 > B1 ***p*** = .020
A1 – D1 n.s.
A1 < NE*
B1 < D1 ***p*** = .023
B1 < NE*
D1 < NE*

A2 (3.86) – B2 (3.77) – NE (2.10)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p < .05$.

Table 5.27. Summary of comparisons carried out on the judges' accent ratings (averaged over consonant segments) with AOL as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each group's mean accent ratings by each judge appear in parentheses.

(a)

Factor: ***onset age of FL learning***
D.V.: Judge 1's FA ratings on consonants
A1 (4.62) – B1 (5.29) – D1 (4.90) – NE (2.35)*
A1 > B1 ***p*** = .036
A1 – D1 n.s.
A1 < NE*
B1 – D1 n.s.
B1 < NE*
D1 < NE*
A2 (5.33) – B2 (5.02) – NE (2.35)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

(b)

Factor: ***onset age of FL learning***
D.V.: Judge 2's FA ratings on consonants
A1 (5.50) – B1 (6.32) – D1 (5.85) – NE (2.52)*
A1 > B1 ***p*** = .022
A1 – D1 n.s.
A1 < NE*
B1 < D1 ***p*** = .072
B1 < NE*
D1 < NE*
A2 (6.17) – B2 (5.33) – NE (2.52)*
A2 < B2 ***p*** = .008
A2 < NE*
B2 < NE*

(c)

Factor: ***onset age of FL learning***
D.V.: Judge 3's FA ratings on consonants
A1 (4.97) – B1 (5.37) – D1 (5.66) – NE (1.97)*
A1 > B1 ***p*** = .097
A1 < D1 ***p*** = .023
A1 < NE*
B1 – D1 n.s.
B1 < NE*
D1 < NE*
A2 (5.80) – B2 (5.92) – NE (1.97)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

(d)

Factor: ***onset age of FL learning***
D.V.: Judge 4's FA ratings on consonants
A1 (4.94) – B1 (4.76) – D1 (4.54) – NE (2.20)*
A1 – B1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – D1 n.s.
B1 < NE*
D1 < NE*
A2 (5.27) – B2 (4.54) – NE (2.20)*
A2 – B2 ***p*** = .028
A2 < NE*
B2 < NE*

Table 5.27 (continued)

(e)

Factor: ***onset age of FL learning*** *D.V.:* Judge 5's FA ratings on consonants
A1 (3.68) – B1 (4.85) – D1 (4.90) – NE (1.73)*
A1 > B1 ***p*** = .005
A1 > D1 ***p*** = .004
A1 < NE*
B1 – D1 n.s.
B1 < NE*
D1 < NE*
A2 (4.87) – B2 (4.34) – NE (1.73)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

(f)

Factor: ***onset age of FL learning*** *D.V.:* Judge 6's FA ratings on consonants
A1 (4.07) – B1 (4.32) – D1 (4.02) – NE (2.51)*
A1 – B1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – D1 n.s.
B1 < NE*
D1 < NE*
A2 (4.22) – B2 (4.19) – NE (2.51)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p < .05$.

5.2.2.2.2. Effect of exposure

5.2.2.2.2.1. Production of vowel segments /i/, /ɪ/, *and* /æ/

In the case of 8-year-old beginners, an increase in exposure did not always result in less accented segment productions – only in /æ/ (Figure 5.40a). Judges 2 and 5 rated A2's vowel production as more foreign-accented than that of A1, while the reverse applied for Judge 3 (Figure 5.41a). In spite of this, no significant effect was found for learners with 200 hours and 416 hours of instruction in English either in their production of English vowels (averaged over judges) or in the six judges' ratings assigned to vowel segments (averaged over vowels) (adjusted $p > .05$).

Likewise, an increase in exposure did not result in significant differences in 11-year-old starters in the ratings assigned by all judges on vowel segments (Figures 5.40b and 5.41b). For the most part, subjects obtained lower FA ratings as their experience in English increased, approaching significance in the ratings assigned by Judge 3 (χ^2 8.889, *df* 2, $p = .012$, adjusted $p > .05$) (see also summary tables 5.28 and 5.29).

In the case of subjects who started to learn the FL at 14 years of age, the ratings obtained for the three vowel segments were less accented when subjects had more exposure. In addition, according to all six judges, more exposure meant less foreign-accented ratings in 14-year-old starters.

Figure 5.40. Mean FA ratings on /i/, /ɪ/ and /æ/ averaged over judges. Factor: exposure to the FL. Results for C1 and C2 are indicative only.

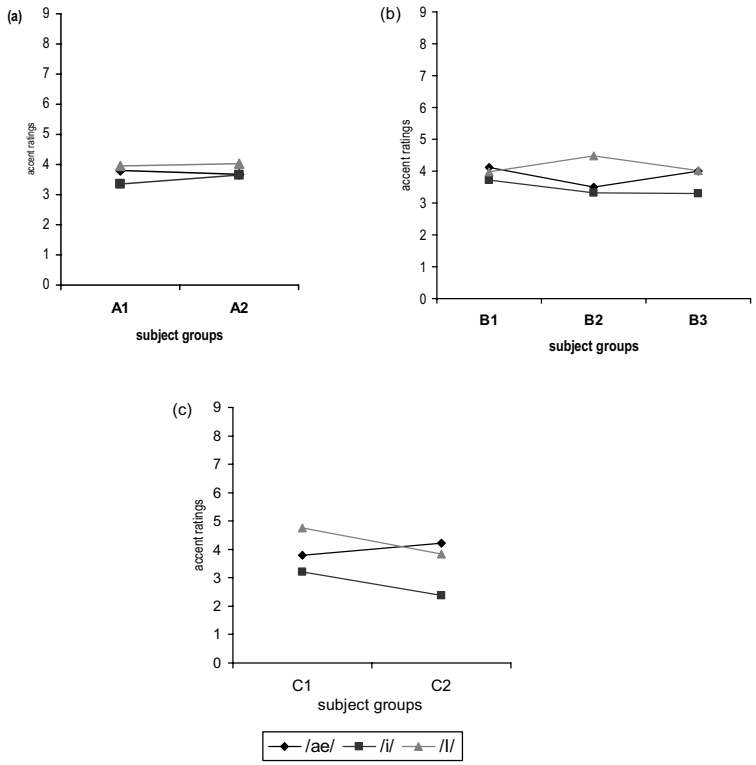
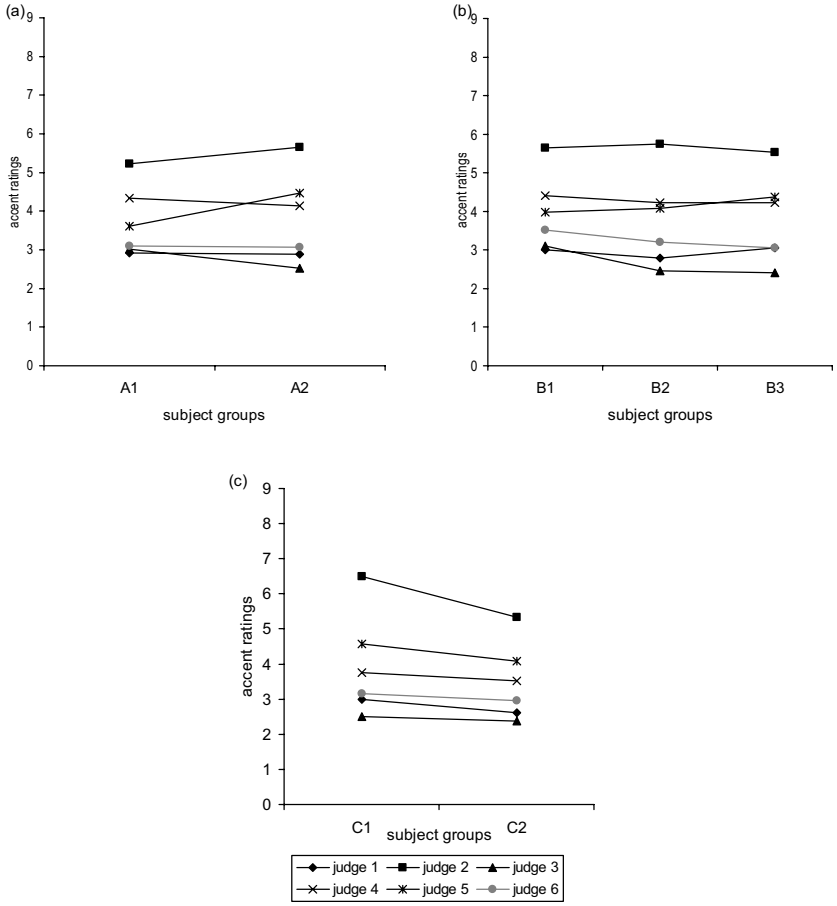


Figure 5.41. Judges' mean FA ratings on /i/, /ɪ/ and /æ/ (averaged over vowel segments). Factor: exposure to the FL. Results for C1 and C2 are indicative only.



As in the evaluation of global FA on words, the degree to which judges considered a segment to be foreign-accented varied greatly from judge to judge (Figure 5.41 above) (see also summary tables 5.28 and 5.29).

Table 5.28. Summary of comparisons carried out on the accent ratings on vowel segments (averaged over judges) with exposure as a factor. Nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as $p =$ .exact significance value. Each group's mean accent ratings for each vowel segment appear in parentheses.

(a)

Factor: ***exposure to the FL*** *D.V.:* FA ratings on /i/
A1 (3.36) – A2 (3.65) n.s.
B1 (3.73) – B2 (3.32) – B3 (3.30) n.s.

(b)

Factor: ***exposure to the FL*** *D.V.:* FA ratings on /ɪ/
A1 (3.96) – A2 (4.02) n.s.
B1 (3.99) – B2 (4.48) – B3 (4.03) n.s.

(c)

Factor: ***exposure to the FL*** *D.V.:* FA ratings on /æ/
A1 (3.80) – A2 (3.67) n.s.
B1 (4.12) – B2 (3.50) – B3 (4.00) ***p*** **=** .054

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p < .05$.

Table 5.29. Summary of comparisons carried out on the judges' accent ratings (averaged over vowels) with exposure to the FL as a factor. Nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each group's mean accent ratings by each judge appear in parentheses.

(a)

Factor: ***exposure to English***
D.V.: Judge 1's FA ratings on vowels
A1 (2.91) – A2 (2.88) n.s.
B1 (3.00) – B2 (2.79) – B3 (3.05) n.s.

(b)

Factor: ***exposure to English***
D.V.: Judge 2's FA ratings on vowels
A1 (5.23) > A2 (5.65) ***p*** = .086
B1 (5.65) – B2 (5.75) – B3 (5.54) n.s.

(c)

Factor: ***exposure to English***
D.V.: Judge 3's FA ratings on vowels
A1 (3.02) < A2 (2.52) ***p*** = .071
B1(3.11) – B2 (2.46) – B3 (2.41) ***p*** = .012

(d)

Factor: ***exposure to English***
D.V.: Judge 4's FA ratings on vowels
A1 (4.34) – A2 (4.13) n.s.
B1 (4.41) – B2 (4.22) – B3 (4.23) n.s.

(e)

Factor: ***exposure to English***
D.V.: Judge 5's FA ratings on vowels
A1 (3.61) > A2 (4.47) ***p*** = .051
B1 (3.98) – B2 (4.08) – B3 (4.37) n.s.

(f)

Factor: ***exposure to English***
D.V.: Judge 6's FA ratings on vowels
A1 (3.10) – A2 (3.07) n.s.
B1 (3.52) – B2 (3.20) – B3 (3.06) n.s.

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p < .05$.

5.2.2.2.2.2. Production of consonant segments /d, s, v/

When FA ratings for each segment averaged over judges were considered separately, no significant differences were found due to the effect of experience ($p > .05$) in 8-year-old beginners, although ratings were higher for A2's /d/s and /s/s than for A1 (Figure 5.42a).

Contrary to the production of vowels, when ratings averaged over consonant segments were examined, the ratings obtained for A1 and A2 differed significantly for judge 3 – and approached significance for judges 1, 2, and 5 – as a result of experience in the TL (*U* 412, *Z* –2.972, $p = .003$, adjusted $p > .05$ for judge 1; *U* 464.5, *Z* –2.403, $p = .016$, adjusted $p > .05$ for judge 2; *U* 370.5, *Z* –3.423, $p = .001$, adjusted $p < .05$ for judge 3; and *U* 396, *Z* –3.135, $p = .002$, $p < .05$ for judge 5). All these judges, as well as judges 4 and 6, rated A2's consonants as more foreign-accented (Figure 5.43a).

As for 11-year-old beginners, Kruskal-Wallis analyses with experience as a factor yielded significant differences in the mean ratings obtained for the three groups on the production of /v/ (χ^2 14.179, *df* 2, $p < .05$). Mann-Whitney *U* tests showed that B1's accent scores for /v/ differed significantly from those of B2 and B3 (*U* 234, *Z* –2.252, $p < .05$; *U* 222.5, *Z* –3.812, $p < .05$). In addition, differences approached significance in /s/ (χ^2 9.179, *df* 2, $p = .010$, adjusted $p > .05$), as well as in the mean FA ratings averaged by consonants for judges 1, 2, and 6 (χ^2 5.751, *df* 2, $p = .056$, adjusted $p > .05$; χ^2 11.694, *df* 2, $p = .003$, adjusted $p > .05$; and χ^2 5.154, *df* 2, $p = .076$, adjusted $p > .05$).

In all these cases, the group with less amount of exposure obtained more foreign-accented scores than the groups with 416 and 726 hours of instruction in the FL. With the exception of B2's production of /d/ and judge 3's ratings for B2, the more exposure subjects had, the less accented their production of consonants was (Figures 5.42b and 5.43b).

By contrast, an increase in exposure often resulted in higher accent scores in 14-year-old starters' with 416 hours of exposure than those with 200 hours of instruction in English (only descriptive level) (Figures 5.42c and 5.43c).

Tables 5.30 and 5.31 summarise the results of the statistical operations performed on the accent ratings on consonant segments with exposure to the FL as a factor.

Figure 5.42. Mean FA ratings on /d/, /s/, and /v/ (averaged over judges). Factor: exposure to the FL. Results for C1 and C2 are indicative only.

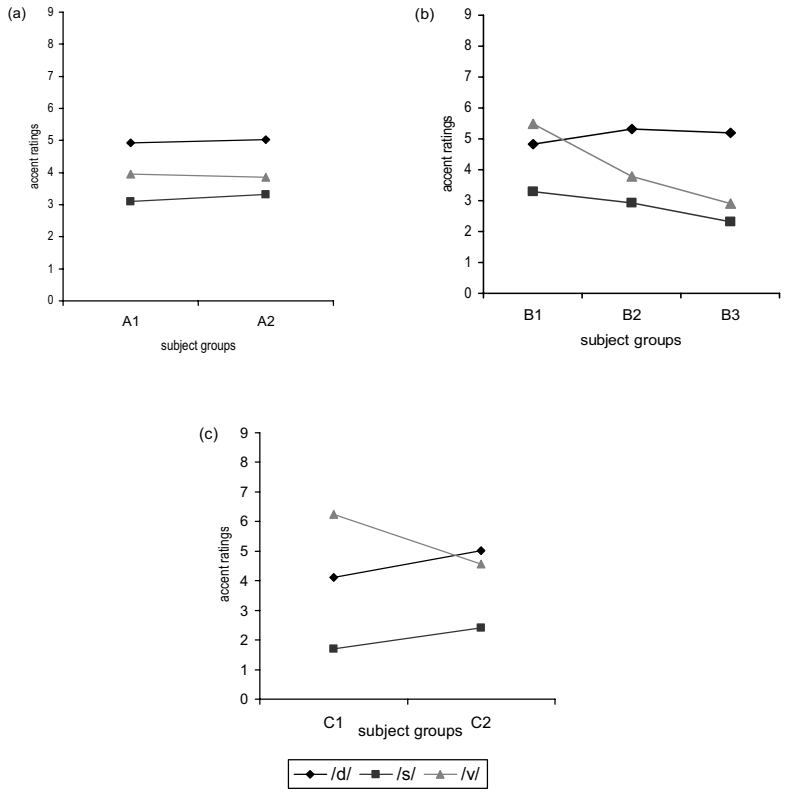


Figure 5.43. Judges' mean FA ratings on /d/, /s/, and /v/ (averaged over consonant segments). Factor: exposure to the FL. Results for C1 and C2 are indicative.

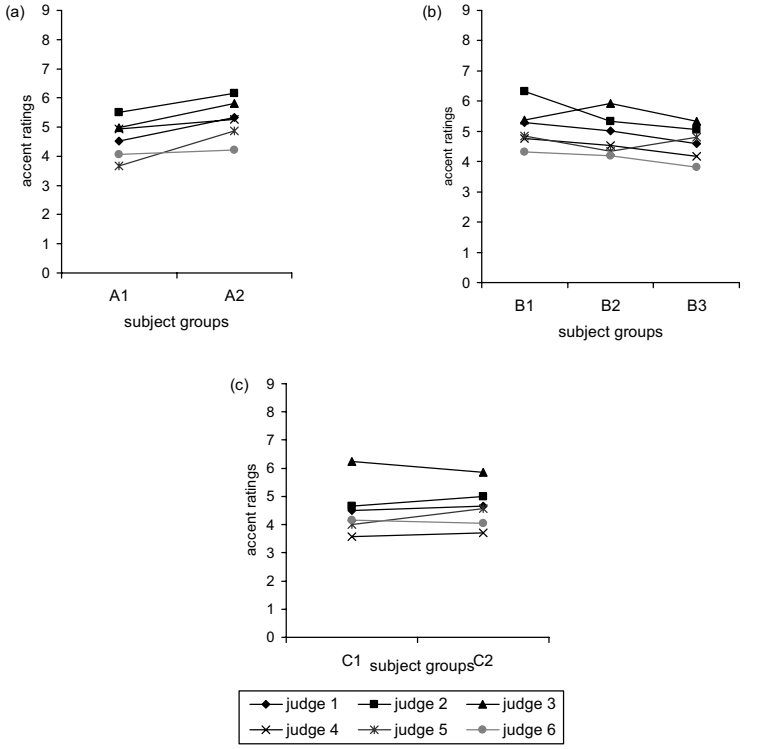


Table 5.30. Summary of comparisons carried out on the accent ratings on consonant segments (averaged over judges) with exposure as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each group's mean accent ratings for each vowel segment appear in parentheses.

(a)

Factor: ***exposure to the FL*** *D.V.:* FA ratings on /d/
A1 (4.92) – A2 (5.03) n.s.
B1 (4.64) – B2 (5.31) – B3 (5.19) n.s.

(b)

Factor: ***exposure to the FL*** *D.V.:* FA ratings on /s/
A1 (3.10) – A2 (3.31) n.s.
B1 (3.29) – B2 (2.93) – B3 (2.32) ***p*** = .010

(c)

Factor: ***exposure to the FL*** *D.V.:* FA ratings on /v/
A1 (3.96) – A2 (3.86) n.s.
B1 (5.50) – B2 (3.77) – B3 (2.91)* B1 – B2 ***p*** = .024 B1 – B3* B2 – B3 n.s.

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p < .05$.

Table 5.31. Summary of comparisons carried out on the judges' accent ratings (averaged over consonants) with exposure to the FL as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each group's mean accent ratings by each judge appear in parentheses.

(a)

Factor: ***exposure to English*** *D.V.:* Judge 1's FA ratings on consonants
A1 (4.62) > A2 (5.33) ***p*** = .003
B1 (5.29) – B2 (5.02) – B3 (4.60) ***p*** = .053

(b)

Factor: ***exposure to English*** *D.V.:* Judge 2's FA ratings on consonants
A1 (5.50) > A2 (6.17) ***p*** = .016
B1 (6.32) – B2 (5.33) – B3 (5.06) ***p*** = .003

Table 5.31. (continued)

(c)

Factor: ***exposure to English*** *D.V.:* Judge 3's FA ratings on consonants
A1 (4.97) > A2 (5.80)*
B1 (5.37) – B2 (5.92) – B3 (5.34) n.s.

(d)

Factor: ***exposure to English*** *D.V.:* Judge 4's FA ratings on consonants
A1 (4.94) – A2 (5.27) n.s.
B1 (4.76) – B2 (4.54) – B3 (4.18) n.s.

(e)

Factor: ***exposure to English*** *D.V.:* Judge 5's FA ratings on consonants
A1 (3.68) > A2 (4.87) ***p*** = .002
B1 (4.85) – B2 (4.34) – B3 (4.81) n.s.

(f)

Factor: ***exposure to English*** *D.V.:* Judge 6's FA ratings on consonants
A1 (4.07) – A2 (4.22) n.s.
B1 (4.32) – B2 (4.19) – B3 (3.82) ***p*** = .076

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p < .05$.

5.2.2.2.3. Effect of dominant L1(s)

5.2.2.2.3.1. Production of vowel segments /i, ɪ, æ/

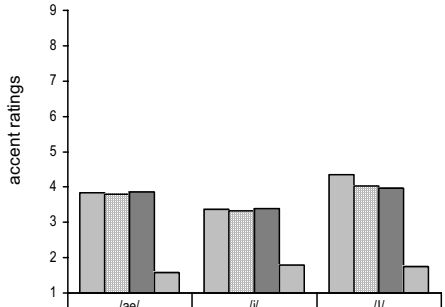
Overall, the three non-native speaker language subgroups' mean FA ratings on vowels, averaged over judges and vowels, were significantly different from subjects whose L1 was English (adjusted $p < .05$) (mean range = 1.19 – 1.93 for NE vs. 2.46 – 5.82 for NNSs).

As regards the three language subgroups (learner groups), there were no significant differences in the ratings obtained for the various groups. Only the average ratings on vowel segments given by judges 2 and 5 to Catalan and Spanish dominant speakers approached significance (U 910.5, Z –2.193, $p < .05$; U 932.5, Z –2.034, $p < .05$). In both cases, Catalan dominant speakers obtained more foreign-accented ratings than Spanish dominant speakers (see Figures 5.44 and 5.45).

When the file was split by subject group, a great deal of variability was encountered concerning which language speaker group obtained lower (or higher) accent scores across all learner groups. Moreover, no language group comparison reached significance – at the most, differences in accent ratings approached significance. More specifically, those differences were located in judge 2's ratings for the three language subgroups of A2 (χ^2 11.728, *df* 2, $p = .003$, adjusted $p > .05$) and judge 4's ratings for B1 (χ^2 7.761, *df* 2, $p = .021$, adjusted $p > .05$).

Summary tables 5.32 and 5.33 below present the results of the statistical operations performed on the accent ratings for vowels with dominant L1(s) as a factor[86].

Figure 5.44. Mean FA ratings on /i/, /ɪ/, and /æ/ averaged over judges. Factor: dominant L1(s).



	/ae/	/i/	/ɪ/
Cat	3.83	3.36	4.34
Sp	3.8	3.33	4.04
Cat/Sp	3.86	3.39	3.97
Eng	1.58	1.79	1.74

[86] As was the case of the accent ratings on words with dominant L1(s) as a factor (5.2.2.1.3), a large number of statistical operations was carried out within each learner group with dominant L1(s) as a factor. Thus, figures and summary tables have not been included in this section, for there were a huge number of different results depending on the learner group and dependent variable being considered. On top of that, none of the comparisons yielded significant results.

Figure 5.45. Judges' mean FA ratings (averaged over vowels). Factor: dominant L1(s).

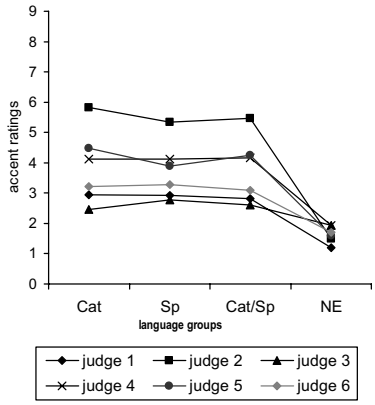


Table 5.32. Summary of comparisons carried out on the accent ratings on vowel segments (averaged over judges) with dominant L1(s) as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). Each language group's mean accent ratings for each vowel segment appear in parentheses.

(a)

Factor: ***dominant L1(s)***   *D.V.:* FA ratings on /i/
Cat (3.36) – Sp (3.33) – C/S (3.39) – NE (1.79)*   Cat – Sp n.s.   Cat – C/S n.s.   Cat < NE*   Sp – C/S n.s.   Sp < NE*   C/S < NE*

(b)

Factor: ***dominant L1(s)***   *D.V.:* FA ratings on /ɪ/
Cat (4.34) – Sp (4.09) – C/S (3.97) – NE (1.74)*   Cat – Sp n.s.   Cat – C/S n.s.   Cat < NE*   Sp – C/S n.s.   Sp < NE*   C/S < NE*

(c)

Factor: ***dominant L1(s)***   *D.V.:* FA ratings on /æ/
Cat (3.83) – Sp (3.80) – C/S (3.86) – NE (1.58)*   Cat – Sp n.s.   Cat – C/S n.s.   Cat < NE*   Sp – C/S n.s.   Sp < NE*   C/S < NE*

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$.

Table 5.33. Summary of comparisons carried out on the judges' accent ratings (averaged over vowel segments) with dominant L1(s) as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each language group's mean accent ratings by each judge appear in parentheses.

(a)

Factor: ***dominant L1(s)***
D.V.: Judge 1's FA ratings on vowels
Cat (2.94) – Sp (2.93) – C/S (2.82) – NE (1.19)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(b)

Factor ***dominant L1(s)***
D.V.: Judge 2's FA ratings on vowels
Cat (5.82) – Sp (5.35) – C/S (5.47) – NE (1.50)*
Cat – Sp ***p*** = .028
Cat – C/S ***p*** = .072
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(c)

Factor: ***dominant L1(s)***
D.V.: Judge 3's FA ratings on vowels
Cat (2.46) – Sp (2.77) – C/S (2.60) – NE (1.93)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(d)

Factor: ***dominant L1(s)***
D.V.: Judge 4's FA ratings on vowels
Cat (4.12) – Sp (4.13) – C/S (4.16) – NE (1.93)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(e)

Factor: ***dominant L1(s)***
D.V.: Judge 5's FA ratings on vowels
Cat (4.48) – Sp (3.90) – C/S (4.24) – NE (1.60)*
Cat – Sp ***p*** = .042
Cat – C/S n.s.
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(f)

Factor: ***dominant L1(s)***
D.V.: Judge 6's FA ratings on vowels
Cat (3.21) – Sp (3.28) – C/S (3.10) – NE (1.70)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p < .05$.

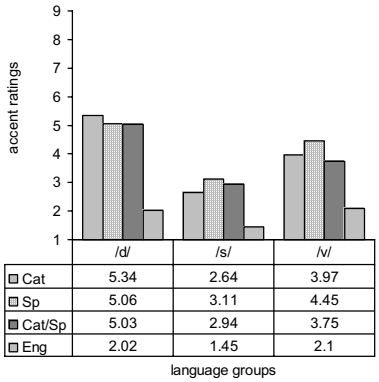
5.2.2.2.3.2. Production of consonant segments /d, s, v/

NE foils' productions of the three consonant segments under study were all rated as native-like by all judges (mean range = 1.45 – 2.52), and, in turn, differed significantly from the three language subgroups (mean range = 2.64 – 5.79), according to Kruskal-Wallis analyses (adjusted $p < .05$).

Among learner groups, mean FA ratings on consonant segments averaged over judges and mean FA ratings for each judge averaged over consonants did not differ significantly from one language subgroup to either of the two remaining language subgroups (Figures 5.46 and 5.47, Tables 5.34 and 5.35).

If each learner group is examined separately, no significant differences were found between the three language subgroups in any of the learner groups. As in the ratings for vowel segments, no specific language group obtained higher or lower accent ratings across all learner groups.

Figure 5.46. Mean FA ratings on /d/, /s/, and /v/ averaged over judges. Factor: dominant L1(s).



	/d/	/s/	/v/
Cat	5.34	2.64	3.97
Sp	5.06	3.11	4.45
Cat/Sp	5.03	2.94	3.75
Eng	2.02	1.45	2.1

Figure 5.47. Judges' mean FA ratings (averaged over consonants). Factor: dominant L1(s).

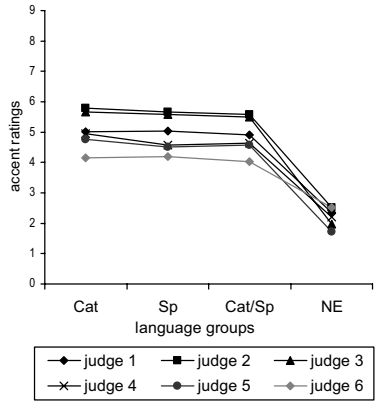


Table 5.34. Summary of comparisons carried out on the accent ratings on consonant segments (averaged over judges) with dominant L1(s) as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). Each language group's mean accent ratings for each consonant segment appear in parentheses.

(a)

Factor: ***dominant L1(s)***   *D.V.:* FA ratings on /d/
Cat (5.34) – Sp (5.06) – C/S (5.03) – NE (2.02)*   Cat – Sp n.s.   Cat – C/S n.s.   Cat < NE*   Sp – C/S n.s.   Sp < NE*   C/S < NE*

(b)

Factor: ***dominant L1(s)***   *D.V.:* FA ratings on /s/
Cat (2.64) – Sp (3.11) – C/S (2.94) – NE (1.45)*   Cat – Sp n.s.   Cat – C/S n.s.   Cat < NE*   Sp – C/S n.s.   Sp < NE*   C/S < NE*

(c)

Factor: ***dominant L1(s)***   *D.V.:* FA ratings on /v/
Cat (3.97) – Sp (4.45) – C/S (3.75) – NE (2.10)*   Cat – Sp n.s.   Cat – C/S n.s.   Cat < NE*   Sp – C/S n.s.   Sp < NE*   C/S < NE*

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$.

Table 5.35. Summary of comparisons carried out on the judges' accent ratings (averaged over consonant segments) with dominant L1(s) as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). Each language group's mean accent ratings by each judge appear in parentheses.

(a)

Factor: ***dominant L1(s)***
D.V.: Judge 1's FA ratings on consonants
Cat (2.94) – Sp (5.02) – C/S (5.04) – NE (2.35)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(b)

Factor ***dominant L1(s)***
D.V.: Judge 2's FA ratings on consonants
Cat (5.82) – Sp (5.79) – C/S (5.67) – NE (2.52)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(c)

Factor: ***dominant L1(s)***
D.V.: Judge 3's FA ratings on consonants
Cat (2.46) – Sp (5.67) – C/S (5.58) – NE (1.97)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(d)

Factor: ***dominant L1(s)***
D.V.: Judge 4's FA ratings on consonants
Cat (4.12) – Sp (4.13) – C/S (4.58) – NE (2.20)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(e)

Factor: ***dominant L1(s)***
D.V.: Judge 5's FA ratings on consonants
Cat (4.48) – Sp (3.90) – C/S (4.51) – NE (1.73)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(f)

Factor: ***dominant L1(s)***
D.V.: Judge 6's FA ratings on consonants
Cat (3.21) – Sp (4.16) – C/S (4.19) – NE (2.51)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat < NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p < .05$.

5.2.2.2.4. Effect of gender

5.2.2.2.4.1. Production of vowel segments /i, ɪ, æ/

Averaged over judges, mean accent ratings on the vowel sounds did not differ significantly between male and female subjects (adjusted $p > .05$). Nor did male and female subjects' ratings (averaged over vowels) differ significantly from each other for any of the six judges (adjusted $p > .05$). Despite the lack of significant differences, female subjects always obtained lower FA ratings than male subjects (see Figures 5.48 and 5.49).

When the file was split by subject group, no gender comparison was statistically significant – though in some cases the differences in accent ratings approached significance. As in the accent ratings obtained for words, female learners received less foreign-accented scores on English vowels than male subjects matched for AOL and experience in the FL.

A summary of the comparisons performed between male and female learners of English on the accent ratings is included in Tables 5.36 and 5.37.

Figure 5.48. Mean FA ratings on /i/, /ɪ/ and /æ/ (averaged over judges). Factor: gender.

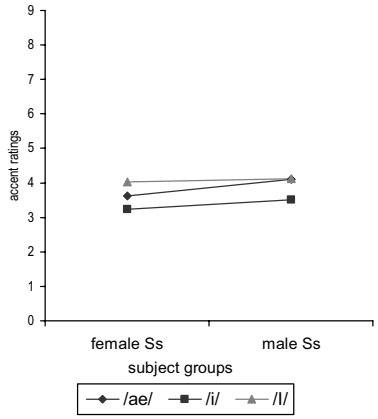


Figure 5.49. Judges' mean FA ratings (averaged over vowel segments). Factor: gender.

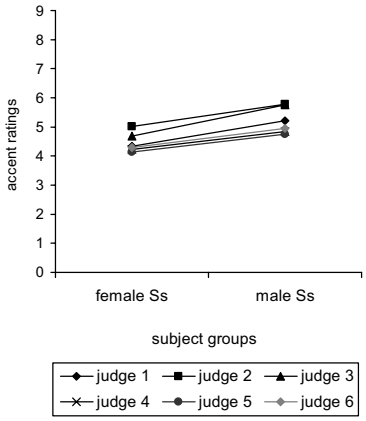


Table 5.36. Summary of comparisons carried out on the accent ratings on vowel segments (averaged over judges) with gender as a factor. Nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as $p =$ .exact significance value. Each gender group's mean accent ratings for each vowel segment appear in parentheses.

(a)

Factor: ***gender*** *D.V.:* FA ratings on /i/
Female (3.24) > Male (3.52) ***p*** = .070

(b)

Factor: ***gender*** *D.V.:* FA ratings on /ɪ/
Female (4.03) – Male (4.12) n.s.

(c)

Factor: ***gender*** *D.V.:* FA ratings on /æ/
Female (3.63) > Male (4.11) ***p*** = .016

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Marginally significant results should be compared to the alpha level of .001 and the resulting adjusted $p <$ .05.

Table 5.37. Summary of comparisons carried out on the judges' accent ratings (averaged over vowel segments) with gender as a factor. Nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each gender group's mean accent ratings by each judge appear in parentheses.

(a)

Factor: ***gender***
D.V.: Judge 1's FA ratings on vowels
Female (2.73) > Male (3.09) ***p*** = .013

(b)

Factor ***gender***
D.V.: Judge 2's FA ratings on vowels
Female (5.29) > Male (5.80) ***p*** = .002

(c)

Factor: ***gender***
D.V.: Judge 3's FA ratings on vowels
Female (2.56) – Male (2.67) n.s.

(d)

Factor: ***gender***
D.V.: Judge 4's FA ratings on vowels
Female (4.02) – Male (4.29) n.s.

(e)

Factor: ***gender***
D.V.: Judge 5's FA ratings on vowels
Female (4.06) – Male (4.39) n.s.

(f)

Factor: ***gender***
D.V.: Judge 6's FA ratings on vowels
Female (3.09) – Male (3.27) n.s.

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p <$ .05.

5.2.2.2.2.4.2. Production of consonant segments /d, s, v/

As was the case of vowel production, female subjects always obtained lower foreign accent scores than male subjects (M = 4.98, 2.74, and 3.82 for female speakers vs. 5.25, 3.08, and 4.16 for male speakers in /d/, /s/, and /v/, respectively). However, gender differences in the production of consonants (averaged over judges) did not yield significant results – only approached significance in /s/ (U 4248.5, Z – 2.792, p = .005, adjusted $p > .05$) (Figure 5.50 and Table 5.38).

Significant differences were found in judges 1's, 3's, and 6's ratings averaged over consonants (adjusted $p < .05$), while for judges 2, 4, and 5 differences in accent ratings between male and female Ss approached significance (Figure 5.51 and Table 5.39).

When ratings were examined separately for each learner group, female subjects' production of English consonant segments was also rated as less foreign-accented than that of male subjects. However, this failed to reveal significant differences between the two gender groups (at the most, they approached significance).

Figure 5.50. Mean FA ratings on /d/, /s/ and /v/ (averaged over judges). Factor: gender.

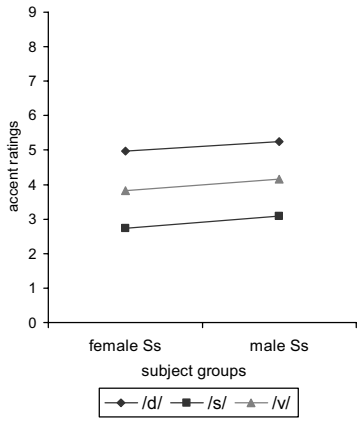


Figure 5.51. Judges' mean FA ratings (averaged over consonant segments). Factor: gender.

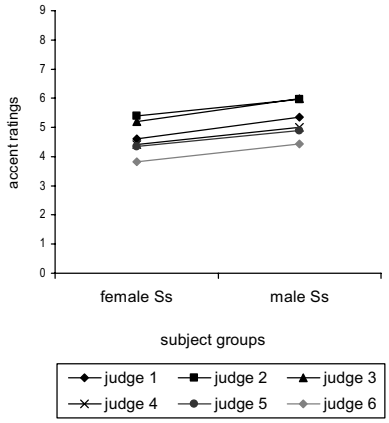


Table 5.38. Summary of comparisons carried out on the accent ratings on consonant segments (averaged over judges) with gender as a factor. Nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each gender group's mean accent ratings for each consonant segment appear in parentheses.

(a)

Factor: ***gender***   *D.V.:* FA ratings on /d/
Female (4.98) – Male (5.25) n.s.

(b)

Factor: ***gender***   *D.V.:* FA ratings on /s/
Female (2.74) > Male (3.08) ***p*** **=** .005

(c)

Factor: ***gender***   *D.V.:* FA ratings on /v/
Female (3.82) – Male (4.16) n.s.

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Marginally significant results should be compared to the alpha level of .001 and the resulting adjusted $p <$ .05.

Table 5.39. Summary of comparisons carried out on the judges' accent ratings (averaged over consonant segments) with gender as a factor. Significant comparisons are marked with * ($p < .05$), *P* values of marginally significant results ($.05 < p < .10$) are stated as $p =$ .exact significance value. Each gender group's mean accent ratings by each judge appear in parentheses.

(a)

Factor: ***gender*** *D.V.:* Judge 1's FA ratings on consonants
Female (4.61) > Male (5.36)*

(b)

Factor ***gender*** *D.V.:* Judge 2's FA ratings on consonants
Female (5.40) > Male (5.97) ***p*** = .002

(c)

Factor: ***gender*** *D.V.:* Judge 3's FA ratings on consonants
Female (5.19) > Male (5.99)*

(d)

Factor: ***gender*** *D.V.:* Judge 4's FA ratings on consonants
Female (4.41) > Male (5.01) ***p*** = .003

(e)

Factor: ***gender*** *D.V.:* Judge 5's FA ratings on consonants
Female (4.34) > Male (4.89) ***p*** = .020

(f)

Factor: ***gender*** *D.V.:* Judge 6's FA ratings on consonants
Female (3.83) > Male (4.43)*

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p < .05$.

5.3.*II.* STUDY 2

The methodology, analyses, and results obtained in Study 2 are described below. As was the case of Study 1, the discussion of results will be presented in Chapter 6.

5.3.1. Method

5.3.1.1. Subjects

The sample of subjects examined in Study 2 was larger than that of Study 1. In other words, data for all learner groups had been collected when Study 2 was conducted. As was done previously (see Sections 4.1 and 5.2.1.1 above), longitudinal subjects were randomly kept either at T1, T2, or T3, so that they appeared only once in their respective group, and thus can be considered cross-sectional subjects for the purpose of analysis. A further criterion was also established in order for a subject to be examined in Study 2: only recordings with a signal-to-noise ratio (S/N) higher than 10.00 dB (see Stimulus Preparation section below) were included in the experiment; hence, the differing group sizes in the AX task, Study 1, and Study 2, and the need to collect more data for the control group. Thus, in the case of Study 2, a total of 161 subjects[87] participated in the experiment, of whom 13 were native speakers of British English.

As for English-speaking subjects, only one subject's recording from Study 1 had a S/N above 10.00 dB. Therefore, more NS data were gathered. The remaining foils were children of British-born residents on the island of Menorca, whose L1 at home was English solely. In fact, those participants made use of languages other than English in the school setting only. They took part in the present experiment on a voluntary basis.

The characteristics of all these participants are summarised in Tables 5.40 and 5.41 (cross-sectional and longitudinal subjects, respectively).

[87] Originally, there were 346 Ss, including both longitudinal and cross-sectional subjects as well as 45 NSs of English.

Table 5.40. Characteristics of participant groups in Study 2 (cross-sectional data). Standard deviations are in parentheses.

Group	N	AOL[a]	Exposure[b]	L1[c]	Gender[d]	Age[e]	Grade[f]
A1	17	8	200	Cat 40% Sp 20% C/S 40%	m 11 f 6	10.97 (.31)	5 *Primaria*
A2	27	8	416	Cat 22.2% Sp 14.8% C/S 63%	m 13 f 14	12.92 (.29)	1 E.S.O.
A3	13	8	726	Cat 23.1% Sp 30.8% C/S 46.2%	m 7 f 6	16.61 (.36)	1*Bachillerato*
B1	13	11	200	Cat 8.3% Sp 58.3% C/S 33.3%	m 5 f 8	13.09 (.43)	7 E.G.B.
B2	14	11	416	Cat 7.1% Sp 35.7% C/S 57.1%	m 8 f 6	14.92 (.30)	1 B.U.P.
B3	15	11	726	Cat 26.7% Sp 26.7% C/S 46.7%	m 10 f 5	18.04 (.27)	C.O.U.
C1	13	14	200	Cat 33.3% Sp 16.7% C/S 50%	m 6 f 6	16.14 (.54)	2 B.U.P.
C2	4	14	416	Cat - Sp 50% C/S 50%	m 2 f 2	18.45 (.86)	C.O.U.
D1	23	18+	200	Cat 26.1% Sp 26.1% C/S 47.8%	m 18 f 5	28.62 (7.9)	2 E.I.
D2	7	18+	416	Cat 28.6% Sp - C/S 71.4%	m 5 f 2	26.25 (5.74)	4 E.I.
D3	2	18+	726	Cat 50% Sp - C/S 50%	m - f 2	45.79 (6.42)	5/6 E.I.
NE	13	—	—	—	m 5 f 8	11.62 (13.98)	—

[a] Onset age of FL learning (in years)
[b] Number of hours of formal exposure to English
[c] Dominant L1(s) (%): Cat (Catalan), Sp (Spanish), C/S (Catalan and Spanish)
[d] m: male, f: female
[e] Mean chronological age at testing (in years)
[f] Subjects' school grade at testing:
Primaria, E.S.O., and *Bachillerato* (new curriculum)
E.G.B., B.U.P., and C.O.U. (former curriculum)
E.I. (Escuela de Idiomas) – (adult) language schools

Table 5.41. Characteristics of longitudinal subjects in Study 2. Standard deviations are in parentheses.

Group	N	AOL[a]	Exposure[b]	L1[c]	Gender[d]	Age[e]	Grade[f]
A1$_{long*}$	5[g]	8	200	Cat 50% Sp - C/S 50%	m 3 f 2	10.83 (.32)	5 *Primaria*
A2$_{long}$	8[g]	8	416	Cat 37.5% Sp 12.5% C/S 50%	m 4 f 4	12.93 (.38)	1 E.S.O.
A3$_{long}$	4[g]	8	726	Cat 75% Sp - C/S 25%	m 2 f 2	16.62 (.39)	1*Bachillerato*
B1$_{long}$	2[h]	11	200	Cat - Sp 50% C/S 50%	m 1 f 1	13.08 (.03)	7 E.G.B.
B2$_{long}$	2[h]	11	416	Cat - Sp 50% C/S 50%	m 1 f 1	15.33 (.00)	1 B.U.P.

*long Longitudinal group
[a] Onset age of FL learning (in years)
[b] Number of hours of formal exposure to English
[c] Dominant L1(s) (%): Cat (Catalan), Sp (Spanish), C/S (Catalan and Spanish)
[d] m: male, f: female
[e] Mean chronological age at testing (in years)
[f] Subjects' school grade at testing:
Primaria, E.S.O., and *Bachillerato* (new curriculum)
E.G.B. and B.U.P. (former curriculum)
[g] Distribution of longitudinal subjects (8-year-old beginners):
4 Ss at Time 1 + Time 2 1 subject at Time 1 + Time 2 + Time 3
3 Ss at Time 2 + Time 3 NO Ss at Time 1 + Time 3
[h] Distribution of longitudinal subjects (11-year-old beginners): 2 Ss at Time 1 + Time 2

5.3.1.2. Task

The task performed was exactly the same as that of Study 1. That is, subjects imitated a list of 34 English words as presented via tape recorder (for a review of the production task, see 4.2). The only difference from the first study was that more data from different groups were added to the present study.

5.3.1.3. Objective

The purpose of Study 2 was to examine whether the variables of onset age of FL learning, amount of exposure to English, dominant L1(s) and gender had any effect on

the production of English vowel sounds by Spanish and/or Catalan learners of English. More precisely, seven English vowel segments were chosen for analysis, namely /i, ɪ, ɛ, æ, ɒ, u, ʌ/. These sounds appeared in a total of eleven words that were part of the 34 words comprising the production task: /i/ in *speak* and *tea*, /ɪ/ in *it* and *this*[88], /ɛ/ in *red* and *tests*, /æ/ in *back* and *pad*, /ɒ/ in *box*, /u/ in *zoo*, and /ʌ/ in *but*[89]. An attempt was made to choose words with the 7 sounds being produced in phonetic contexts that were not thought to have influenced their production to the learners' disadvantage – mainly unvoiced plosive context (but see, e.g., Cebrian, 2002c; García Lecumberri, 1999). Furthermore, the vowel segments selected looked at sounds that are often examined in L2 phonological acquisition research, /i, ɪ, ɛ, æ/, and other segments that are not so frequently investigated, /ɒ, u, ʌ/, in Spanish and Catalan learners of English (e.g. Flege et al., 1994; Rallo, 2005; for Italian learners of English, see, among others, Munro et al., 1996).

Two parts made up Study 2, i.e. a FA rating task and a vowel identification task. The first part was an extension of the accent rating task carried out on vowel and consonant segments in Study 1. In this case, the number of learner groups was increased. The second part was designed to describe subjects' production of English vowels, in addition to their possible mispronunciations. It was expected that this would provide a more complete picture of FL learners' production of the TL vowels.

5.3.1.4. Stimulus preparation

Prior to the final stimulus preparation, S/N was measured on the low back vowel /ɒ/ in *box* with the software Signalyze version 3.0 (Signal Analysis for Speech and Sound) (Keller, 1994). This procedure was carried out in order to have a more homogeneous set of data and to (perhaps) help obtain a higher degree of inter-judge

[88] /ɪ/ in *this* was the only sound that had been examined in the same word in the previous study.

[89] It would have been desirable to examine /ɒ/, /u/, and /ʌ/ in two instances, like /i/, /ɪ/, /ɛ/, and /æ/. However, the production task did not contain any more words with the vowels /ɒ/ and /u/. In the case of /ʌ/, the word *bumped* was not included in the rating of /ʌ/, since subjects had a great deal of difficulty perceiving/understanding this word, and thus they mispronounced the entire word from the very beginning (as noted down by the two raters in the pilot study, as well as by the author's impressionistic perception on digitising the data).

agreement. In other words, it was hypothesised that by having recordings with a smaller range of higher S/N, the apparent effect of very noisy recordings on judges' ratings might be diminished. The measurement of the signal was made at the vowel midpoint, while that of noise was taken in the interval between the end of the production of the word as delivered by the taped model voice and before the beginning of the Ss' production of the word *box* (approximately 200 ms before the Ss' beginning of word production). Then, all these values and ratios were entered and calculated with the statistical package SPSS 11.0 for Windows. A total of 346 measurements on /ɒ/ (and the corresponding "noisy" intervals) was made: 281 productions belonging to learners and 45 to English foils. S/N ranged from –.36 dB to 24.13 dB. Looking back at Study 1, the recordings examined had a S/N that spanned across the whole range obtained, of which 27.5% of the recordings was below 5 dB. Therefore, one could hypothesise that listeners in Study 1 might have had a great deal of difficulty in rating word and segment productions (with a S/N level below 5 dB) for degree of FA, since studies such as Oyama (1978/1982) have reported that English NSs rendered sentences presented in masking white noise with a S/N level of 5 dB as unintelligible. Moreover, the available findings and observations suggest that there might be a cut-off point between S/N levels of 5 dB and 10 dB where speech is still intelligible. For instance, Munro (1998) estimated that a mean S/N of 7.9 dB would result in native English listeners' consideration of both English and Mandarin NSs' production of English sentences as partially or completely unintelligible. Additionally, Kent (1997) has noted that "in a typical classroom the teacher's speech is about 6 dB more intense than the background noise. This is a marginal S/N for effective speech communication." (p. 395). Finally, Flege and Liu (2001) showed that non-native subjects performed on a similar basis when English stimuli were presented in ideal (no-noise) conditions and in differing S/N levels – i.e. 10 dB and 16 dB. Based on all of the above, it was decided that for the present experiment (Study 2) only those recordings with a S/N higher than 10.00 dB would be selected for further study[90]. That meant that between 22.9% and 61.5% of the recordings for each subject group were discarded for Study 2 – averaged over groups, 51.2% were left out – resulting in the 161 Ss of this experiment.

[90] Preliminary analyses on the data from Study 1 were further undertaken in order to test the hypothesis that listeners would disagree at higher rates over noisier recordings than less noisy ones. The intraclass coefficients obtained did not confirm this assumption for the most part (with the exception of the ratings on the segments /i/, /s/, and /v/ – .63, .89, and .92 vs. .62, .78, and .87 for recordings with a S/N above and below 10.00 dB, respectively). In spite of this, the cut-off point of 10.00 dB was still applied in order to prevent any possible (negative) effects of a wider range of S/N recordings on the listeners' ratings, such as listener fatigue and therefore less reliable judgements.

The 1,754 imitations obtained (161 Ss x 11 words[91]) were digitised with CoolEdit at 22.05 kHz and 16-bit resolution. Next, they were normalised to 75% for peak intensity.

Different random blocks were created for each sound presentation in the specific words and for each of the 7 judges, both for the FA rating task and for the vowel identification task, amounting to a total of 154 different randomised blocks (11 words x 7 judges x 2 tasks).

As in Study 1, this experiment was set up and conducted by means of WNSPARCS (Smith, 1997).

5.3.1.5. Listeners

Seven female NSs of General Canadian English took part in Study 2 as paid judges. None of the listeners in the present study had participated in Study 1. For the present experiment, listeners were required to have "a good ear" and skills for phonetic transcription. Thus, they were either undergraduate or graduate students in Linguistics at the University of Ottawa, all of whom had taken courses in phonetics. Their mean age was 26.14 years (range = 21 – 40 years) and reported normal hearing. All judges except judge 6 had some familiarity with Spanish, as they had taken courses in that language at some point. Their age of first use or exposure was between 7 and 21 years, although one listener had actually spent two months in a Latin American Spanish-speaking country. No judge was familiar with Catalan or had spent any time in a Catalan-speaking area. As for French, all listeners had studied French and had lived in a community in which French is often used. However, just one judge spoke French on a daily basis. Finally, one judge knew some Italian and American Sign Language, another listener knew some Japanese, and two reported some knowledge of German.

5.3.1.6. Procedure

The task that listeners performed consisted of two parts, amounting to a total of 22 sessions per judge. The first eleven sessions involved assigning FA ratings to each of the 7 vowels distributed in 11 words on a 9-point scale of FA (like Study 1, *1* meant *no*

[91] As noted earlier, subjects missed repeating a word on several occasions. For this study, the total number of imitations produced per word was the following: 161 productions for *back*, *box*, *red*, *speak*, and *zoo* (thus, no missing cases), 160 productions for *but*, *it*, *tea*, and *tests* (1 missing case each), 157 imitations for *this* (4 missing cases), and 152 imitations for *pad* (9 missing cases).

FA, *5* a *moderate amount of FA*, and *9* was used to indicate *very strong FA*). At the beginning of every block, judges read a set of specific instructions where they were asked to assign a rating to each vowel produced by the 161 Ss and to ignore background noise as much as possible (see Appendix D for the instructions given for both parts 1 and 2). They were also informed that all blocks contained both NS and NNS vowel productions. In addition, 5 practice items of the taped model voice were also included at the beginning, so that listeners became familiar with the model that the Ss had imitated, as well as with the software and rating technique. Listeners' responses to those items did not count. Both a written stimulus and audio file were presented at the same time on the computer screen and via headphones, respectively. Listeners then had 1.5 seconds to assign a rating by clicking on one of the buttons of the scale on the computer screen, but they could, if they wished, listen to an item as many times as needed.

The remaining 11 sessions consisted of identifying the vowel sound that the subjects had produced. One vowel (in one word) was presented at a time. Moreover, judges were provided with a table with 15 response options on the computer screen, from which they had to choose the option that best characterised the vowel segment in question. The first three buttons in the first row stood for three degrees of the intended vowel: good instance, slightly distorted, and very distorted vowel production. The remaining buttons/rows provided judges with possible mispronunciations – the second and third rows, in particular, contained the most likely mispronunciations[92].

Specifically, the possible response options for /i/ were: "[ij][93] good", "[ij] slightly distorted", "[ij] very distorted", "[i]", "[e]", "[ej]", "[ɪ]", "[ɛ]", "[æ]", "[ɑ]", "[ɒ]", "[ʌ]", "[u]", and "[ɜˑ] or [əˑ]".

For /ɪ/, the options were: "[ɪ] good", "[ɪ] slightly distorted", "[ɪ] very distorted", "[i]", "[ij]", "[ɛ]", "[e]", "[ej]", "[æ]", "[ɑ]", "[ɒ]", "[ʌ]", "[u]", and "[ɜˑ] or [əˑ]".

As for /ɛ/, listeners could choose among: "[ɛ] good", "[ɛ] slightly distorted", "[ɛ] very distorted", "[i]", "[ij]", "[ɪ]", "[e]" "[ej]", "[æ]", "[ɑ]", "[ɒ]", "[ʌ]", "[ʊ]", "[u]", and "[ɜˑ] or [əˑ]".

---

[92] Most options were based on the two raters' assessments in the pilot study, the author's observations on digitising the subject data, and findings of previous studies (e.g. Flege, 1991a).

[93] [j] in [ij] and [w] in [uw] referred to the fact that these English vowels in question – /i/ and /u/ – are diphthongised and different from the pure tense, non-diphthongised, vowels, as would be the Spanish and/or Catalan [i]- and [u]-quality vowels. In addition, [j] in [ej] and [w] in [ow] indicated that the vowels preceding them are diphthongised, as well.

In the case of /æ/, the responses available were: "[æ] good", "[æ] slightly distorted", "[æ] very distorted", "[a]", "[ɑ]", "[ɒ]", "[ɔ]", "[ɛ]", "[e]", "[ej]", "[ʌ]", "[ʊ]", "[u]", "[i]", and "[ɝ] or [ɚ]".

As far as /ɒ/ was concerned, responses could be: "[ɒ] good", "[ɒ] slightly distorted", "[ɒ] very distorted", "[ɑ]", "[ɔ]","[ʌ]", "[ʊ]", "[u]", "[ow]", "[uw]", "[i]", "[æ]", "[a]", "[ɛ]", and "[ɝ] or [ɚ]".

For /u/, the options were: "[uw] good", "[uw] slightly distorted", "[uw] very distorted", "[u]", "[ʊ]", "[ɔ]", "[ow]", "[ʌ]", "[ɒ]", "[ɑ]", "[æ]", "[a]", "[ɛ]", "[i]", and "[ɝ] or [ɚ]".

Finally, the possible response options for /ʌ/ were: "[ʌ] good", "[ʌ] slightly distorted", "[ʌ] very distorted", "[æ]", "[a]", "[ɑ]", "[ɒ]", "[ɔ]", "[ʊ]", "[u]", "[ɛ]", "[e]", "[ow]", "[i]", and "[ɝ] or [ɚ]".

For the actual visual arrangement of all the response options for each vowel segment examined, see Appendix E.

When it came to studying the judges' identifications, "good", "slightly distorted", and "very distorted" realisations of each segment were considered correct identifications, for they involved recognition of the target sound, despite its degree of distortion. Therefore, they were grouped into one response for subsequent statistical analyses. The remaining vowel responses were considered misidentifications.

Like the FA rating task, judges read a list of specific instructions for each vowel segment at the beginning of each session. In this case, they were provided with 10 practice items of a male NS of General Canadian English who produced both the intended vowel categories and some possible mispronunciations. Listeners' responses to those items did not count. Nor were they given feedback, partly because it was feared that they would become discouraged in case they did not identify all the pronunciations correctly. The inter-stimulus interval was 1.5 seconds and they could listen to an item as many times as required.

Moreover, both in the FA rating and vowel identification tasks a random 25% of the total subjects' productions was added immediately after the 161 word imitations had been presented. In order to prevent listeners from losing concentration due to lengthy blocks, in those cases where a vowel sound was looked at in two words, only one word contained the extra 25% repeated productions – about 40 word productions. Thus,

unknown to the listeners, the sessions with /i, ɪ, ɛ, æ, ɒ, u, ʌ/ in *tea*, *it*, *red*, *back*, *box*, *zoo*, and *but*, respectively, presented the added 25%. As a result, 14 sessions contained 201 stimuli (7 blocks for the FA rating task and another 7 for the vowel identification task) and the remaining 4 blocks had 161 stimuli.

Each session lasted 30–35 minutes. Listeners did a maximum of 4 sessions per day, each one separated by at least a 10-minute break. Blocks belonging to different parts were carried out on different days. The experiment took 154 sessions (22 blocks x 7 listeners) – about a month and a half – to be completed. A total of 14,357 FA ratings ([201 vowel productions x 7 blocks x 7 judges] + [161 vowel productions x 4 blocks x 7 judges]) was obtained, as well as 14,357 vowel identifications.

SPSS 11.0 for Windows was used to analyse listeners' accent ratings and vowel identifications.

5.3.2. Results

5.3.2.1. Intra-rater reliability

As stated in 5.3.1.6. Procedure, for the study of intra-listener consistency the sessions examining /i, ɪ, ɛ, æ, ɒ, u, ʌ/ in the words *tea*, *it*, *red*, *back*, *box*, *zoo*, and *but* contained a random extra set of repeated subjects' productions both for the FA rating task and the vowel identification task. The extra set of repeated productions was always administered immediately after all the 161 subjects' productions had been delivered within the same session. Therefore, judges were unaware of the fact that they would rate a number of instances twice. The set of repeated productions was the same for a given segment, but the order of presentation was counterbalanced across the seven listeners. Moreover, those sets varied – i.e. they did not contain the same subjects' productions – in the seven segments under study. It should also be noted that, although the repeated productions were chosen at random, in the randomisation procedure it was ensured that at least one subject from each group was included in the sets of repeated samples. In all sessions the extra set amounted to a total of 25% of the total subjects' productions.

In order to determine intra-judge reliability (or intra-rater/intra-listener reliability), intra-class correlation coefficients were computed for each judge, on the one hand, between the accent ratings assigned in the first presentation of the repeated

productions and those assigned in the second presentation; and, on the other hand, between the vowel identifications obtained in the first and second presentations.

As far as the FA rating task is concerned, the intra-class correlation coefficients obtained indicated that for the most part an acceptable degree of intra-rater agreement was achieved (coefficients between .70 and .80). Moreover, on only five occasions were the reliability coefficients below .60: .53, .47, and .54 for judges 2, 3, and 4 in /ɛ/; and .47 and .45 for judges 3 and 5 in /ʌ/. All the coefficients obtained for each judge and vowel sound are presented in Table 5.42, where it can be seen that judges were fairly consistent in assigning the same accent ratings to those subjects' productions that were heard twice.

Table 5.42. Judges' intra-class correlation coefficients for FA ratings.

	/i/	**/ɪ/**	**/ɛ/**	**/æ/**	**/ɒ/**	**/u/**	**/ʌ/**
Judge 1	.89	.82	.73	.80	.77	.75	.85
Judge 2	.69	.73	.53	.65	.66	.62	.70
Judge 3	.78	.79	.47	.63	.72	.71	.47
Judge 4	.88	.86	.54	.61	.85	.82	.85
Judge 5	.74	.83	.71	.60	.83	.77	.45
Judge 6	.67	.75	.63	.64	.73	.69	.78
Judge 7	.80	.82	.74	.68	.83	.72	.78

As in Southwood and Flege (1999), percent agreement scores between the two presentations were calculated for each judge's ratings, in addition to percent scores of those ratings that differed by one point (above and below) on the scale of FA between the first and second presentations[94]. These results are summarised in Table 5.43, where it can be observed that percent agreement scores ranged from 9.8% to 73.3% and percent +/–1 scores ranged from 8.9% to 55.5%. These results also corroborate the finding of high and acceptable degrees of intra-listener agreement, since listeners were consistent in the

[94] Both one-point difference scores above (+1 score) and below (–1 score) the original accent ratings were averaged into one score, for it was not expected that listeners would consistently rate the repeated samples more foreign-accented (or less foreign-accented) as is the case of studies where listeners are provided with training/feedback on specific traits that might characterise foreign-accented speech before rating the same samples for a second time. In this study, the effect of familiarity with subjects' FA speech on the ratings judges assigned did not come into play, because the repeated productions were included within the same session with no explicit training beforehand.

ratings assigned, and when their ratings differed from one presentation to the other, it was mostly by one point on the scale.

Table 5.43. Judges' percent agreement and +/–1 scores for accent ratings.

	/i/		/ɪ/		/ɛ/		/æ/		/ɒ/		/u/		/ʌ/	
Judges	=	+/–1	=	+/–1	=	+/–1	=	+/–1	=	+/–1	=	+/–1	=	+/–1
1	31.1	26.7	38.9	33.3	27.3	27.3	20.5	33.3	28.9	24.5	20.5	41.1	22	34.2
2	48.9	37.8	50	11.1	60.6	27.3	35.9	46.1	53.3	28.9	38.5	28.2	65.9	21.9
3	28.9	26.7	19.4	36.1	30.3	36.3	20.5	18	31.1	20	20.5	28.2	9.8	24.4
4	42.2	35.5	27.8	47.2	15.2	30.3	12.8	30.8	26.7	42.2	28.2	28.2	36.6	31.8
5	60	17.8	36.1	22.2	51.5	18.2	20.5	33.3	33.3	28.9	38.5	17.9	19.5	34.1
6	22.2	28.9	38.9	19.5	27.3	27.3	17.9	33.3	35.6	31.2	25.6	28.2	19.5	36.6
7	44.4	55.5	52.8	19.5	51.5	27.2	33.3	15.4	73.3	8.9	30.8	35.9	46.3	24.4

Last, it is worth noting that a high degree of intra-listener agreement does not imply a high degree of inter-rater agreement (at least, a necessary degree to be able to pool judges' ratings into one single rating, as reported in 5.3.2.2 below). In other words, judges were fairly consistent within themselves, but while one judge might have rated vowel productions as consistently less foreign-accented, another listener might have rated those vowel productions as more foreign-accented.

As for the vowel identification task, the intra-class reliability coefficients obtained were mostly acceptable, like the FA rating task (see Table 5.44). However, unlike the accent rating task, the reliability coefficients were more diverse in that there were more instances of both high and low degree of intra-judge reliability (coefficients above .80 and below .70 for high and low degree, respectively).

Table 5.44. Judges' intra-class correlation coefficients for vowel identifications.

	/i/	/ɪ/	/ɛ/	/æ/	/ɒ/	/u/	/ʌ/
Judge 1	.83	.75	.65	.65	.61	.78	.71
Judge 2	.66	.85	.99	.89	.62	.76	.69
Judge 3	.45	.75	.99	.44	.59	.73	.35
Judge 4	.83	.82	.99	.19	.86	.82	.37
Judge 5	.99	.83	.72	.86	.71	.95	.84
Judge 6	–.09	.62	.49	.88	.76	.60	.76
Judge 7	.80	.69	.87	.20	.82	.75	.78

Percent agreement scores between the first and second vowel identifications obtained for the same subjects were also calculated. In that case, the percent agreement scores included both the intended target sound identified as such in the two presentations and the vowel sounds other than the target that were identified as the same segment in both presentations. Thus, in Table 5.45 it can be seen that, in general, the agreement percentages obtained for vowel identifications were higher than those of accent scores (compare with Table 5.43 above). This is probably due to the fact that the vowel identification task mostly presented vowels as specific response options, whereas the FA rating task involved somehow subjective responses in that each judge had to decide by themselves what constituted a higher or lower degree of FA.

The percent agreement scores also illustrate the finding that the judges' vowel identifications were reliable. Averaged over judges, the agreement scores were as follows: 80.9% for /i/, 61.3% for /ɪ/, 88.4% for /ɛ/, 65.4% for /æ/, 66.5% for /ɒ/, 76% for /u/, and 68.1% for /ʌ/. It should also be mentioned that those instances where judges showed discrepancies within themselves involved mainly vowels that are close in their acoustic phonetic space to the intended target sound. So, in the disagreements found it was often the case that the intended target sound was identified in either the first or second presentation, and a close vowel in the phonetic space was then identified in the other presentation. More precisely, /æ/ was misidentified with [a] and [ɑ]; /ʌ/ with [a] and [æ]; /ɒ/ with [ɑ][95], [ɔ] and [ʌ]; and /ɛ/ with [e]. Unexpectedly, there seemed to be some degree of subjectivity in the responses for /i/ (and by extension /ɪ/) and /u/, as often English /i/ and /u/ were misidentified as non-diphthongised tense vowels [i] and [u] in either the first or second presentation. It could be hypothesised that the number of choices facing the raters (listeners), namely slightly distorted, very distorted, non-diphthongised, as well as a range of other vowel choices, created a set of distinctions that is too fine-grained and too subjective to have resulted in a higher rate of inter-listener agreement.

[95] In addition, /ɒ/-/ɑ/ are not distinguished in American English.

Table 5.45. Judges' percent agreement scores for vowel identifications. Main disagreements involving the identification of the correct target sound in either presentation and a different sound are also stated in the column "dis".

	/i/		/ɪ/		/ɛ/		/æ/		/ɒ/		/u/		/ʌ/	
Judges	=	dis	=	dis	=	dis	=	dis	=	dis	=	dis	=	dis
1	57.2	16.6	53.7	12.2	76.2	19	44.8	31	47.4	29	77.5	17.5	48.5	31.4
2	95.3	4.8	73.2	17	92.9	7.1	72.4	20.6	78.9	21	77.5	17.5	65.7	22.9
3	90.5	9.5	56.1	22	92.9	7.1	65.5	31	57.9	31.6	62.5	25	62.9	25.7
4	85.7	11.9	56.1	21.9	97.6	2.4	58.6	34.4	73.7	7.9	80	17.5	62.8	34.2
5	92.9	7.1	63.4	9.8	95.3	4.8	93.1	6.8	55.3	26.3	90	10	71.4	20
6	66.6	33.3	51.2	34.2	71.5	21.4	55.1	31	78.9	21.1	67.5	25	85.7	14.3
7	78.6	16.7	75.6	29.5	92.8	7.2	68.9	27.5	73.7	18.4	77.5	17.5	80	11.4

Finally, it would have been desirable to carry out correlations between each judge's accent ratings and vowel identifications. However, this could not be accomplished for a series of reasons. First, the correlations were not computed on the entire dataset because there was low inter-listener agreement, which in turn would have involved calculating correlations for each judge and vowel sound separately. Consequently, it would have been difficult to reach clear-cut conclusions, based on the large number of coefficients obtained. Second, although the results would have been more manageable with a reduced dataset (such as the intra-judge reliability results reported above), it was not possible to carry out this analysis, for the extra repeated samples were not identical in the FA rating task and in the vowel identification task.

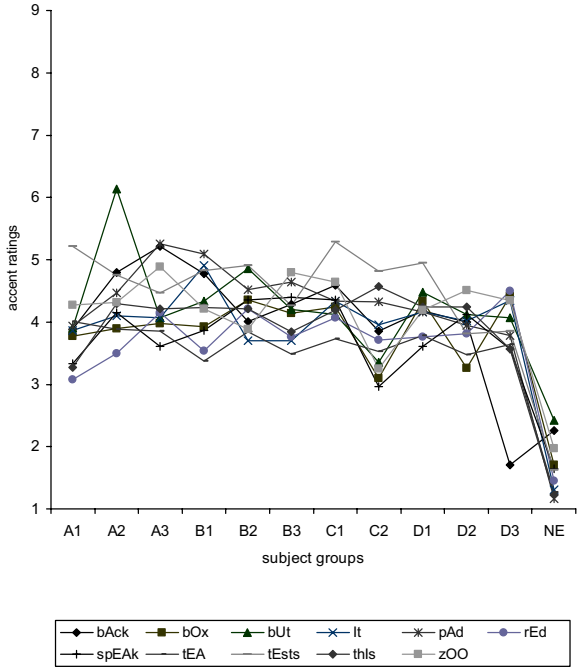
5.3.2.2. FA ratings on vowels /i, ɪ, ɛ, æ, ɒ, u, ʌ/

Overall, judges used all the points on the scale of FA when rating the seven vowel segments. The sole exception was judge 2, who on occasion seemed reluctant to use points 7–9 on the scale.

Mean accent ratings on each of the vowel sounds averaged over judges were calculated for each of the 161 Ss in the study. Each judge's FA ratings averaged over vowels (11 words) for each subject were also computed. These results are displayed in Figures 5.52 and 5.53 as a function of subject group. In those figures, it can be seen that

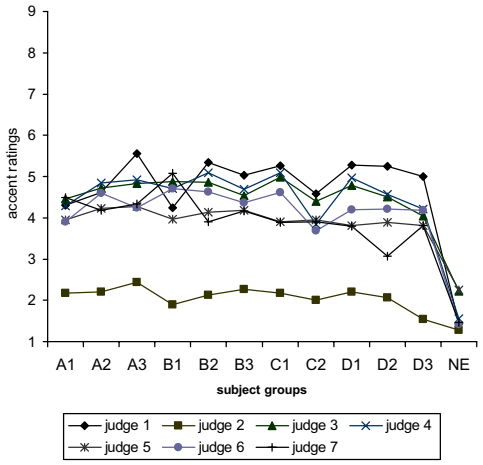
English foils' productions were judged as more native-like than FL learners' (mean range = 1.16 – 2.42 for NE group vs. 1.9 – 6.14 for learners[96]).

Figure 5.52. Mean FA ratings on vowels (averaged over 7 judges).



[96] Results for C2, D2, and D3 are presented as indicative of what ratings might have been obtained, had they been larger groups. As seen in Table 5.40, very few subjects comprised those groups. Consequently, these groups' performance was not analysed statistically.

Figure 5.53. Judges' mean FA ratings (averaged over vowels).



Averaged over the 161 Ss, the mean accent ratings obtained for each vowel segment (in each word) were as follows: 3.54 and 3.74 for /i/ in *tea* and *speak*; 3.86 and 3.85 for /ɪ/ in *it* and *this*; 3.55 and 4.52 for /ɛ/ in *red* and *tests*; 4.21 and 4.19 for /æ/ in *back* and *pad*; 3.83 for /ɒ/ in *box*; 4.16 for /u/ in *zoo*; and 4.07 for /ʌ/ in *but*. Reliability coefficients on averaged accent scores for each vowel across 7 judges were computed. Cronbach's alpha was .82. At first sight, accent ratings on vowels were therefore fairly similar (range = 3.54 – 4.52, only a one-point difference on the scale). A repeated measures ANOVA was performed on the vowel foreign accent ratings (11 levels) as repeated measures and with age of FL learning (5 levels) and experience in English (4 levels) as factors. The simple main effect for vowel was significant (F (10, 126) = 3.467, $p < .001$). No two-way interaction or three-way interaction was significant ($p > .05$). Homogeneity tests were also significant in 5 vowels ($p < .05$): /i/ in *tea*, /ɪ/ in *it* and *this*, /æ/ in *pad*, /ɒ/ in *box*, and /ʌ/ in *but*. Pairwise comparisons revealed that /æ/ in *pad* was significantly more foreign-accented than /ɛ/ in *red* and /i/ in *tea* (Bonferroni $p < .05$). Two divergent results were found for /ɛ/ depending on the word in which it appeared: /ɛ/ in *red* was rated as less foreign-accented than /æ/ in *pad* and /ɛ/ in *tests*, whereas /ɛ/ in

tests was significantly more foreign-accented than /i/ in *speak* and *tea* (Bonferroni $p <$ .05).

Averaged over 161 Ss, mean vowel accent ratings for each judge were the following: 4.66, 2.03, 4.44, 4.40, 3.86, 4.06, and 3.84 for judges 1, 2, 3, 4, 5, 6, and 7. Inter-listener reliability was examined on the 7 judges' accent ratings, yielding a Cronbach's alpha value of .93. Next, the accent ratings were submitted to a (5) age of FL learning x (4) exposure x (7) listener repeated measures ANOVA with listener as a repeated measure. The simple main effect for listener was significant (F (6, 130) = 167.843, $p < .001$). And so were the listener x exposure two-way interaction (F (12, 260) = 2.287, $p < .001$) and listener x age of FL learning x exposure three-way interaction (F (30, 522) = 1.766, $p < .001$). The listener x age of FL learning two-way interaction was nonsignificant (F (18, 368.181) = 1.462, $p = .100$). Homogeneity tests were significant in judge 2's ratings ($p < .05$). Pairwise comparisons showed that judge 2 rated vowel productions as significantly less accented than the remaining six judges (Bonferroni $p <$ .05). Besides, judges 3 and 4 rated subjects' vowels as significantly more foreign-accented than judges 2, 5, 6, and 7 (Bonferroni $p < .05$).

The results reported recommended using nonparametric tests, as well as looking at each vowel sound and judge's ratings separately. Due to the large number of tests with the same subject-group comparisons, the alpha level was set at .001 to maintain an experiment-wise error of .05 (.001 x 7 judges x 7 vowels). In addition, and prior to statistical analyses, a further attempt was made in order to reduce data – namely vowels /i/, /ɪ/, /ɛ/ and /æ/ that were studied in two words. Wilcoxon Signed Ranks Tests revealed that ratings on /i/ in *speak* and *tea*, /ɪ/ in *it* and *this*, and /æ/ in *back* and *pad* were not significantly different (Z –.180, $p = .857$ for /i/; Z –.304, $p = .761$ for /ɪ/; and Z –1.477, $p = .140$ for /æ/); whereas accent ratings on /ɛ/ in *red* and *tests* did differ from each other significantly (Z –6.648, $p < .001$). Therefore, the nonparametric test coincided with the parametric pairwise comparisons performed in the repeated measures ANOVA mentioned above in that only the FA ratings on /ɛ/ in two words were significantly different (Bonferroni $p < .05$). Thus, /ɛ/ in *red* was rated as less foreign-accented than /ɛ/ in *tests*. Consequently, ratings on /i/, /ɪ/, and /æ/ were each pooled into a single rating, but not those of /ɛ/.

5.3.2.2.2. Effect of onset age of FL learning

Kruskal-Wallis analyses on the 7 judges' average vowel ratings as dependent variables and with onset age of FL learning as a factor showed that there were significant differences in all judges' ratings among groups ($p < .05$). All listeners assessed the control group's vowels as significantly less foreign-accented than those of the learner groups, according to Mann-Whitney *U* tests ($p < .05$).

Age differences among learners did not result in significantly higher or lower ratings for a given group, with the sole exception of B1–D1 in judge 7's ratings (*U* 37, *Z* –3.426, adjusted $p < .05$). At the most, differences in ratings approached significance.

Thus, in the various age groups with 200 hours of instruction the age comparisons that were close to being significant were as follows: A1–B1 in judge 6's ratings (*U* 67, *Z* –1.822, $p = .069$, adjusted $p > .05$); A1–C1 in judges 1's and 6's ratings (*U* 35.5, *Z* –2.068, $p = .039$, adjusted $p > .05$; *U* 40, *Z* –1.815, $p = .070$, adjusted $p > .05$); A1–D1 in judges 1's, 4's, and 7's ratings (*U* 99, *Z* –2.278, $p = .023$, adjusted $p > .05$; *U* 112.5, *Z* –1.879, $p = .060$, adjusted $p > .05$; *U* 114, *Z* –1.834, $p = .067$, adjusted $p > .05$); B1–C1 in judges 1's, 2's, and 7's ratings (*U* 24, *Z* –2.140, $p = .032$, adjusted $p > .05$; *U* 27.5, *Z* –1.897, $p = .058$, adjusted $p > .05$; *U* 15, *Z* –2.778, $p = .005$, adjusted $p > .05$); and B1–D1 in judges 1's, 2's. and 6's ratings (*U* 61, *Z* –2.560, $p = .010$, adjusted $p > .05$; *U* 70, *Z* –2.245, $p = .025$, adjusted $p > .05$; *U* 86, *Z* –1.660, $p = .097$, adjusted $p > .05$). Judges differed in terms of which age group obtained lower (or higher) ratings. For instance, in the comparisons involving 8-year-old starters – A1–B1, A1–C1, and A1–D1 – judges 1, 4, and 6 rated their English vowel productions as less foreign-accented than older starters, whereas judge 7 rated adult learners' (D1) productions as less foreign-accented than 8-year-old beginners' vowel productions. Another example of the variability in judges' ratings can be found in the comparison B1–C1: while B1 was rated as less foreign-accented than C1 by judge 1, judge 7 rated B1 as more foreign-accented than C1 (see Figure 5.54a).

With 416 hours of instruction, age differences in ratings between A2 and B2 only approached significance in judge 1's evaluations (*U* 94.5, *Z* –1.801, $p = .072$, adjusted $p > .05$), in favour of 8-year-old beginners (Figure 5.54b). Moreover, the same variability among judges' ratings as that of learners with 200 hours of instruction was observed.

Last, with 726 hours of exposure to the TL, the ratings judges assigned to the different age groups were not significantly different. Nor did they approach significance (Figure 5.54c). In this case, the tendency observed was for 8-year-old beginners to produce vowels with a slightly higher degree of FA than 11-year-old starters.

Table 5.46 below summarises the results of the statistical operations performed on the judges' ratings with AOL as a factor.

Figure 5.54. Judges' ratings on vowels. Factor: age of FL learning. Results for C2, D2, and D3 are indicative only.

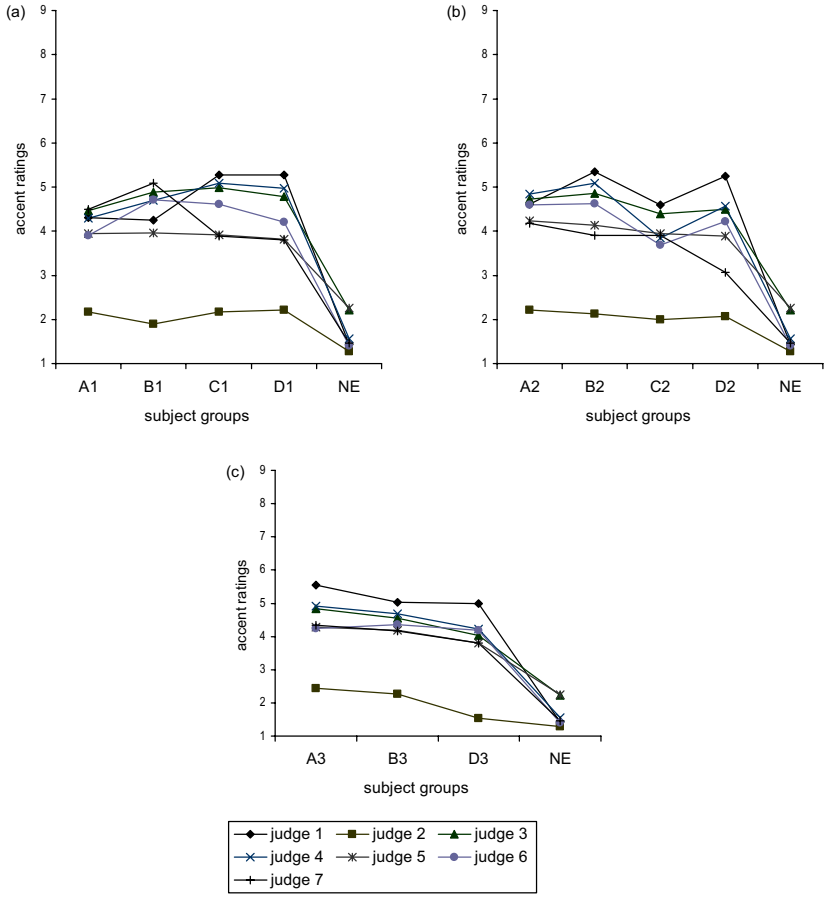


Table 5.46. Summary of comparisons carried out on the judges' accent ratings (averaged over vowels) with AOL as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each group's mean accent ratings by each judge appear in parentheses.

(a)

Factor: ***onset age of FL learning***
D.V.: Judge 1's FA ratings

A1 (4.30) – B1 (4.25) – C1 (5.27) – D1 (5.28) – NE (1.43)*
A1 – B1 n.s.
A1 > C1 ***p*** = .039
A1 > D1 ***p*** = .023
A1 < NE*
B1 > C1 ***p*** = .032
B1 > D1 ***p*** = .010
B1 < NE*
C1 – D1 n.s.
C1 < NE*
D1 < NE*

A2 (4.61) – B2 (5.34) – NE (1.43)*
A2 > B2 ***p*** = .072
A2 < NE*
B2 < NE*

A3 (5.55) – B3 (5.03) – NE (1.43)*
A3 – B3 n.s.
A3 < NE*
B3 < NE*

(b)

Factor: ***onset age of FL learning***
D.V.: Judge 2's FA ratings

A1 (2.17) – B1 (1.90) – C1 (2.17) – D1 (2.21) – NE (1.28)*
A1 – B1 n.s.
A1 – C1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 > C1 ***p*** = .058
B1 > D1 ***p*** = .025
B1 < NE*
C1 – D1 n.s.
C1 < NE*
D1 < NE*

A2 (2.21) – B2 (2.12) – NE (1.28)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

A3 (2.43) – B3 (2.27) – NE (1.28)*
A3 – B3 n.s.
A3 < NE*
B3 < NE*

(c)

Factor: ***onset age of FL learning***
D.V.: Judge 3's FA ratings

A1 (4.46) – B1 (4.88) – C1 (4.98) – D1 (4.78) – NE (2.22)*
A1 – B1 n.s.
A1 – C1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – C1 n.s.
B1 – D1 n.s.
B1 < NE*
C1 – D1 n.s.
C1 < NE*
D1 < NE*

A2 (4.72) – B2 (4.86) – NE (2.22)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

A3 (4.83) – B3 (4.54) – NE (2.22)*
A3 – B3 n.s.
A3 < NE*
B3 < NE*

(d)

Factor: ***onset age of FL learning***
D.V.: Judge 4's FA ratings

A1 (4.29) – B1 (4.70) – C1 (5.09) – D1 (4.97) – NE (1.56)*
A1 – B1 n.s.
A1 – C1 n.s.
A1 > D1 ***p*** = .060
A1 < NE*
B1 – C1 n.s.
B1 – D1 n.s.
B1 < NE*
C1 – D1 n.s.
C1 < NE*
D1 < NE*

A2 (4.84) – B2 (5.09) – NE (1.56)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

A3 (4.92) – B3 (4.69) – NE (1.56)*
A3 – B3 n.s.
A3 < NE*
B3 < NE*

Table 5.46 (continued)

(e)

Factor: ***onset age of FL learning***   *D.V.:* Judge 5's FA ratings
A1 (3.95) – B1 (3.96) – C1 (3.91) – D1 (3.81) – NE (2.25)*   A1 – B1 n.s.   A1 – C1 n.s.   A1 – D1 n.s.   A1 < NE*   B1 – C1 n.s.   B1 – D1 n.s.   B1 < NE*   C1 – D1 n.s.   C1 < NE*   D1 < NE*
A2 (4.23) – B2 (4.13) – NE (2.25)*   A2 – B2 n.s.   A2 < NE*   B2 < NE*
A3 (4.28) – B3 (4.18) – NE (2.25)*   A3 – B3 n.s.   A3 < NE*   B3 < NE*

(f)

Factor: ***onset age of FL learning***   *D.V.:* Judge 6's FA ratings
A1 (3.90) – B1 (4.71) – C1 (4.61) – D1 (4.20) – NE (1.41)*   A1 > B1 ***p*** = .081   A1 > C1 ***p*** = .070   A1 – D1 n.s.   A1 < NE*   B1 – C1 n.s.   B1 < D1 ***p*** = .097   B1 < NE*   C1 – D1 n.s.   C1 < NE*   D1 < NE*
A2 (4.60) – B2 (4.63) – NE (1.41)*   A2 – B2 n.s.   A2 < NE*   B2 < NE*
A3 (4.25) – B3 (4.36) – NE (1.41)*   A3 – B3 n.s.   A3 < NE*   B3 < NE*

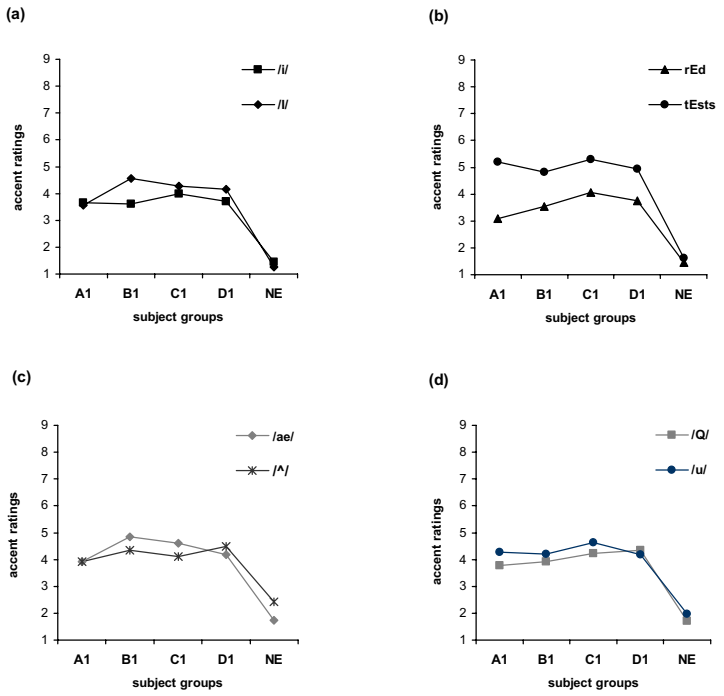
(g)

Factor: ***onset age of FL learning***   *D.V.:* Judge 7's FA ratings
A1 (4.49) – B1 (5.08) – C1 (3.89) – D1 (3.80) – NE (1.46)*   A1 – B1 n.s.   A1 – C1 n.s.   A1 < D1 ***p*** = .067   A1 < NE*   B1 < C1 ***p*** = .005   B1 – D1 n.s.   B1 < NE*   C1 < D1*   C1 < NE*   D1 < NE*
A2 (4.18) – B2 (3.90) – NE (1.46)*   A2 – B2 n.s.   A2 < NE*   B2 < NE*
A3 (4.34) – B3 (4.17) – NE (1.46)*   A3 – B3 n.s.   A3 < NE*   B3 < NE*

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p < .05$.

Further Kruskal-Wallis analyses were carried out on the mean FA scores for each vowel (averaged over the 7 judges) and with age of FL learning as a factor, all yielding significant differences among groups (adjusted $p < .05$). Mann-Whitney U tests showed that NE subjects' production of the 7 vowels was rated as significantly less foreign-accented than that of learners ($p < .05$) (range = 1.27 – 2.42 for the control group vs. 2.75 – 5.29 for learner groups), except for /ʌ/ in the following comparisons: A1–NE (U 40, Z –2.954, $p = .003$, adjusted $p > .05$), B1–NE (U 25.5, Z –3.028, $p = .002$, adjusted $p > .05$), and B3–NE (U 31, Z –3.068, $p = .002$, adjusted $p > .05$). Despite this, the differences between those learner groups and the NE group were nearly significant.

Figure 5.55. FA ratings on /i, ɪ, ɛ, æ, ɒ, u, ʌ/. Factor: onset age of FL learning. Note: /Q/ = /ɒ/



Among subjects with different starting ages of FL learning and with 200 hours of instruction in English, no between-group comparison yielded significant results ($p > .05$ in all Mann-Whitney *U* tests performed). However, A1 tended to receive lower accent ratings on /ɪ/, /ɛ/ (*red*), /æ/, /ɒ/ and /ʌ/ than did the remaining age groups matched for exposure. As seen in Figure 5.55 above, these learners obtained ratings of the point 3 on the scale (range = 3.08 – 3.92) – i.e. some degree of FA, but less than a moderate amount of FA. The most noticeable exceptions to this tendency were the ratings obtained for /ɛ/ (*tests*) (5.21) and /u/ (4.25). For /i/ and /ɛ/ (*tests*), B1 obtained lower ratings, while for /u/ it was D1 that received lower ratings. The largest difference observed was a one-point difference on the scale between A1 and B1 (3.57 and 4.57, respectively) in the production of /ɪ/. This difference was not large enough to be significant, though it approached significance (*U* 67, *Z* –1.822, $p = .069$, adjusted $p > .05$).

When learners' exposure amounted to 416 hours, age differences were not so clear-cut (see Figure 5.56 below). Depending on the segment assessed, either A2 or B2 obtained lower FA ratings[97]. Ratings were somehow higher in learners with 416 hours of instruction than with 200 hours. A striking case was that of /ʌ/, where A2 received an average accent score of 6.14. By contrast, B2's /ʌ/ received a mean rating of 4.86. The difference in the two groups' ratings was only close to being significant (*U* 118.5, *Z* –1.943, $p = .052$, adjusted $p > .05$).

Last, accent ratings on the 7 vowels were more similar across the various age groups with 726 hours of exposure than across the other age group comparisons with different degrees of exposure reported above. But there was variability as to which age group obtained lower FA ratings depending on the vowel segment being examined, as Figure 5.57 below shows. Mann-Whitney *U* tests revealed that no between-group comparison on every vowel segment reached significance. Moreover, it was observed that ratings for A3 and B3 were often more foreign-accented than those of A1 and B1, respectively.

[97] Recall that ratings of groups C2 and D2 are only indicative. If more Ss had made up those groups, and if the same ratings had been obtained, then an advantage of late starters over early starters would have been observed.

Figure 5.56. FA ratings on /i, ɪ, ɛ, æ, ɒ, u, ʌ/. Factor: onset age of FL learning. Note: /Q/ = /ɒ/. Results for C2 and D2 are indicative only.

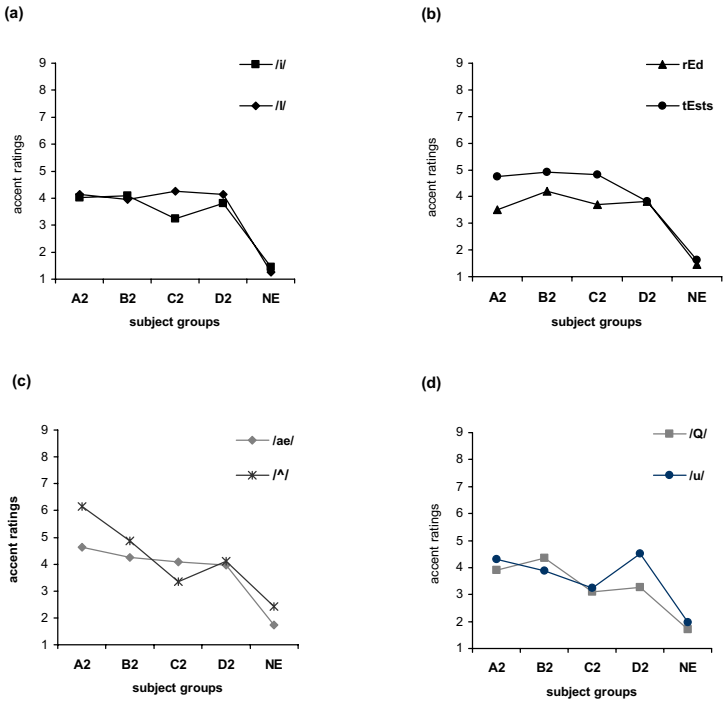
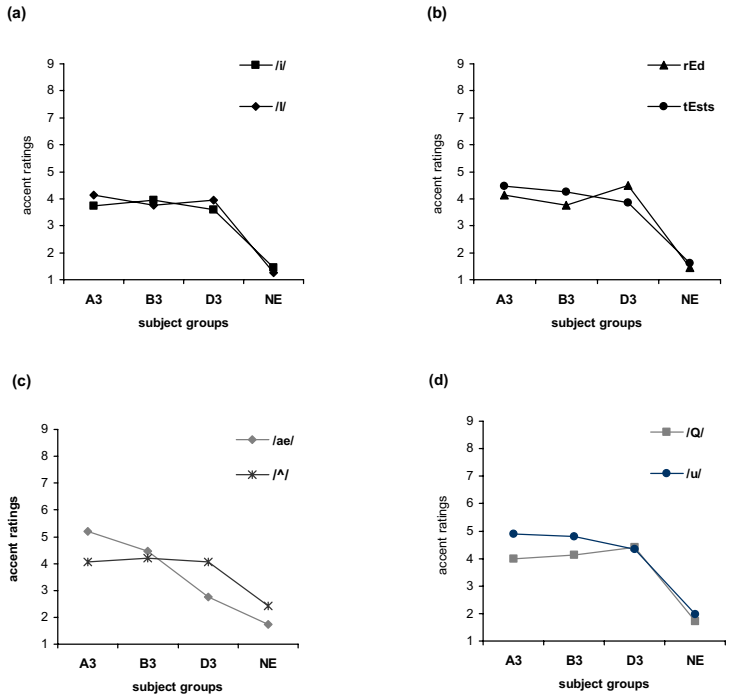


Figure 5.57. FA ratings on /i, ɪ, ɛ, æ, ɒ, u, ʌ/. Factor: onset age of FL learning. Note: /Q/ = /ɒ/. Results for D3 are indicative only.



Next, Table 5.47 summarises the statistical results obtained for each of the vowel sounds examined with onset age of FL learning as a factor.

Table 5.47. Summary of comparisons carried out on the accent ratings on vowels (averaged over judges) with AOL as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each group's mean accent ratings for each vowel appear in parentheses.

(a)

Factor: ***onset age of FL learning***
D.V.: FA ratings on /i/

A1 (3.67) – B1 (3.62) – C1 (3.98) – D1 (3.70) – NE (1.46)*
A1 – B1 n.s.
A1 – C1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – C1 n.s.
B1 – D1 n.s.
B1 < NE*
C1 – D1 n.s.
C1 < NE*
D1 < NE*

A2 (4.02) – B2 (4.10) – NE (1.46)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

A3 (3.73) – B3 (3.95) – NE (1.46)*
A3 – B3 n.s.
A3 < NE*
B3 < NE*

(b)

Factor: ***onset age of FL learning***
D.V.: FA ratings on /ɪ/

A1 (3.57) – B1 (4.57) – C1 (4.27) – D1 (4.16) – NE (1.27)*
A1 > B1 ***p*** = .069
A1 – C1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – C1 n.s.
B1 – D1 n.s.
B1 < NE*
C1 – D1 n.s.
C1 < NE*
D1 < NE*

A2 (4.13) – B2 (3.96) – NE (1.27)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

A3 (4.14) – B3 (3.77) – NE (1.27)*
A3 – B3 n.s.
A3 < NE*
B3 < NE*

(c)

Factor: ***onset age of FL learning***
D.V.: FA ratings on /ɛ/ *red*

A1 (3.08) – B1 (3.54) – C1 (4.07) – D1 (3.76) – NE (1.45)*
A1 – B1 n.s.
A1 > C1 ***p*** = .057
A1 > D1 ***p*** = .075
A1 < NE*
B1 – C1 n.s.
B1 – D1 n.s.
B1 < NE*
C1 – D1 n.s.
C1 < NE*
D1 < NE*

A2 (3.50) – B2 (4.21) – NE (1.45)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

A3 (4.15) – B3 (3.76) – NE (1.45)*
A3 – B3 n.s.
A3 < NE*
B3 < NE*

(d)

Factor: ***onset age of FL learning***
D.V.: FA ratings on /ɛ/ *tests*

A1 (5.21) – B1 (4.83) – C1 (5.29) – D1 (4.95) – NE (1.62)*
A1 – B1 n.s.
A1 – C1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – C1 n.s.
B1 – D1 n.s.
B1 < NE*
C1 – D1 n.s.
C1 < NE*
D1 < NE*

A2 (4.75) – B2 (4.91) – NE (1.62)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

A3 (4.47) – B3 (4.26) – NE (1.62)*
A3 – B3 n.s.
A3 < NE*
B3 < NE*

Table 5.47 (continued)

(e)

Factor: ***onset age of FL learning***
D.V.: FA ratings on /æ/

A1 (3.92) – B1 (4.85) – C1 (4.61) – D1 (4.17) – NE (1.73)*
A1 – B1 n.s.
A1 – C1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – C1 n.s.
B1 < D1 ***p*** = .083
B1 < NE*
C1 – D1 n.s.
C1 < NE*
D1 < NE*

A2 (4.64) – B2 (4.26) – NE (1.73)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

A3 (5.21) – B3 (4.46) – NE (1.73)*
A3 – B3 n.s.
A3 < NE*
B3 < NE*

(f)

Factor: ***onset age of FL learning***
D.V.: FA ratings on /ɒ/

A1 (3.77) – B1 (3.93) – C1 (4.24) – D1 (4.34) – NE (1.71)*
A1 – B1 n.s.
A1 – C1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – C1 n.s.
B1 – D1 n.s.
B1 < NE*
C1 – D1 n.s.
C1 < NE*
D1 < NE*

A2 (3.90) – B2 (4.36) – NE (1.71)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

A3 (3.98) – B3 (4.14) – NE (1.71)*
A3 – B3 n.s.
A3 < NE*
B3 < NE*

(g)

Factor: ***onset age of FL learning***
D.V.: FA ratings on /u/

A1 (4.27) – B1 (4.21) – C1 (4.64) – D1 (4.19) – NE (1.97)*
A1 – B1 n.s.
A1 – C1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – C1 n.s.
B1 – D1 n.s.
B1 < NE*
C1 – D1 n.s.
C1 < NE*
D1 < NE*

A2 (4.31) – B2 (3.89) – NE (1.97)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

A3 (4.89) – B3 (4.80) – NE (1.97)*
A3 – B3 n.s.
A3 < NE*
B3 < NE*

(h)

Factor: ***onset age of FL learning***
D.V.: FA ratings on /ʌ/

A1 (3.91) – B1 (4.34) – C1 (4.12) – D1 (4.48) – NE (2.42)*
A1 – B1 n.s.
A1 – C1 n.s.
A1 – D1 n.s.
A1 < NE ***p*** = .003
B1 – C1 n.s.
B1 – D1 n.s.
B1 < NE ***p*** = .002
C1 – D1 n.s.
C1 < NE*
D1 < NE*

A2 (4.10) – B2 (4.86) – NE (2.42)*
A2 – B2 n.s.
A2 < NE ***p*** = .052
B2 < NE*

A3 (4.07) – B3 (4.20) – NE (2.42)*
A3 – B3 n.s.
A3 < NE*
B3 < NE*

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p < .05$.

5.3.2.2.2. Effect of exposure

Contrary to what might be expected, judges tended to rate 8-year-old beginners' production of English vowels as more foreign-accented, as their experience in the TL increased. This tendency was more noticeable in learners with 416 hours of instruction in English – that is, A2 often received higher FA scores than A1 and A3 (see Figure 5.58a below). None of the differences in ratings between the three groups was significant, though in the cases of judges 1 and 6 differences in ratings approached significance (χ^2 7.292, *df* 2, p = .026, adjusted $p > .05$; χ^2 5.734, *df* 2, p = .057, adjusted $p > .05$). In this case, the ratings assigned to learners with 200 hours of exposure in English (M = 4.3 and 3.9 for judges 1 and 6, respectively) were lower than those of learners with 416 hours (M = 4.61 and 4.60) and with 726 hours of instruction (M = 5.55 and 4.25).

In relation to 11-year-old beginners, judges differed as to which group they assigned lower accent scores as a result of an increase in exposure to English. Thus, while judges 1, 2, 4, and 5 assigned more foreign-accented ratings to the groups with a higher amount of exposure, judges 3, 6, and 7 rated groups with more exposure as less foreign-accented than groups with less instruction in English (Figure 5.58b). The accent scores that judges 1 and 7 assigned to 11-year-old starters approached significance (χ^2 6.029, *df* 2, p = .049, adjusted $p > .05$; χ^2, 6.940, *df* 2, p = .031, adjusted $p > .05$).

In general terms, an increase in exposure in older learners (groups C and D) led to rather less foreign-accented scores. Thus, in spite of judges' variability, they agreed in assigning lower accent scores to both 14-year-old and adult starters as the amount of instruction in English increased. As mentioned above, due to the small number of subjects that formed some of these groups, the results were not analysed statistically.

Table 5.48 below summarises the results of the statistical analyses conducted on ratings of groups A and B.

Figure 5.58. Judges' ratings on vowels. Factor: exposure to the FL.

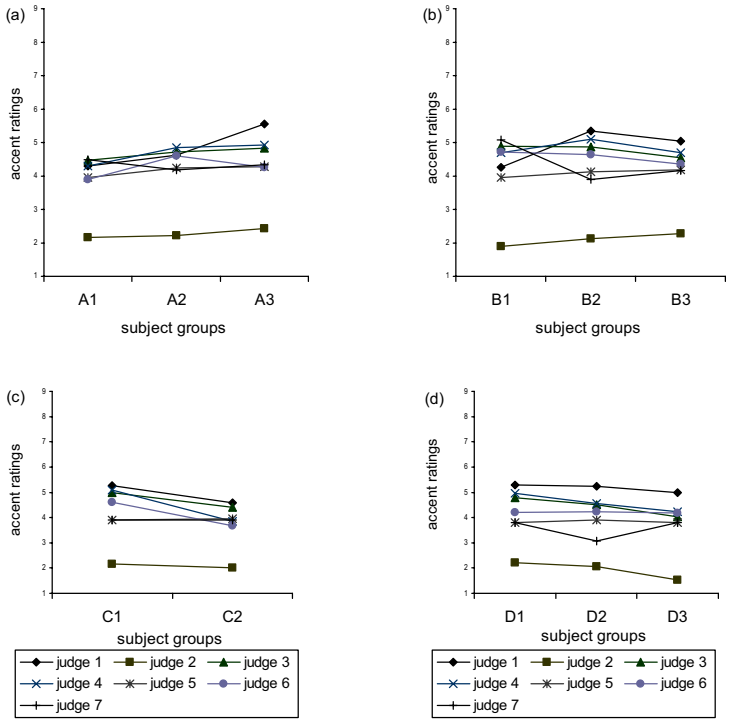


Table 5.48. Summary of comparisons carried out on the judges' accent ratings (averaged over vowels) with exposure as a factor. Nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as $p =$ .exact significance value. Each group's mean accent ratings by each judge appear in parentheses.

(a)

Factor: ***exposure to English*** *D.V.:* Judge 1's FA ratings
A1 (4.30) – A2 (4.61) – A3 (5.55) ***p*** = .026
B1 (4.25) – B2 (5.34) – B3 (5.03) ***p*** = .049

(b)

Factor: ***exposure to English*** *D.V.:* Judge 2's FA ratings
A1 (2.17) – A2 (2.21) – A3 (2.43) n.s.
B1 (1.90) – B2 (2.12) – B3 (2.27) n.s.

(c)

Factor: ***exposure to English*** *D.V.:* Judge 3's FA ratings
A1 (4.46) – A2 (4.72) – A3 (4.83) n.s.
B1 (4.88) – B2 (4.86) – B3 (4.54) n.s.

(d)

Factor: ***exposure to English*** *D.V.:* Judge 4's FA ratings
A1 (4.29) – A2 (4.84) – A3 (4.92) n.s.
B1 (4.70) – B2 (5.09) – B3 (4.69) n.s.

(e)

Factor: ***exposure to English*** *D.V.:* Judge 5's FA ratings
A1 (3.95) – A2 (4.23) – A3 (4.28) n.s.
B1 (3.96) – B2 (4.13) – B3 (4.18) n.s.

(f)

Factor: ***exposure to English*** *D.V.:* Judge 6's FA ratings
A1 (3.90) – A2 (4.60) – A3 (4.25) ***p*** = .057
B1 (4.71) – B2 (4.63) – B3 (4.36) n.s.

(g)

Factor: ***exposure to English*** *D.V.:* Judge 7's FA ratings
A1 (4.49) – A2 (4.18) – A3 (4.34) n.s.
B1 (5.08) – B2 (3.90) – B3 (4.17) ***p*** = .031

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p < .05$.

FA scores on vowels (averaged over judges) were also submitted to separate Kruskal-Wallis analyses with exposure to the FL as a factor to assess its effect on the subjects' production of English vowel segments.

Figure 5.59 shows the FA ratings that group A (8-year-old starters) with varying degrees of exposure received. A first look at this figure points to the fact that an increase in exposure did not lead to less foreign-accented vowel productions. Rather, learners produced the English vowels with a higher amount of FA, as their exposure to English increased. The only exception was /ɛ/ (*tests*), which showed the expected effect of exposure – namely, the more exposure to the TL Ss had, the lower the accent ratings were. Out of the differences in ratings observed between groups, only the difference in ratings on /ɛ/ (*red*) and /æ/ between the three groups approached significance (χ^2 5.289, *df* 2, $p = .071$, adjusted $p > .05$; χ^2 5.436, *df* 2, $p = .066$, adjusted $p > .05$).

11-year-old beginners also presented a similar pattern to that of 8-year-old beginners: learners with more hours of instruction in English received higher scores. By the same token, it was often the case that subjects halfway through their learning (according to the research design) were the group that obtained the most accented scores (see Figure 5.60). Nevertheless, Kruskal-Wallis analyses did not result in any significant difference between the three learner groups in the ratings on the seven English vowels examined (adjusted $p > .05$).

Furthermore, it was not possible to determine whether FA ratings were higher or lower for new or similar sounds. In other words, there was a great deal of variability as to which type of sounds – new or similar – consistently received lower (or higher for the like) accent ratings in either age group (8- and 11-year-old beginners) at each specific point of instruction (200 hours, 416 hours, and 726 hours).

Figure 5.59. FA ratings on group A's production of /i, ɪ, ɛ, æ, ɒ, u, ʌ/. Factor: exposure to the FL. Note: /Q/ = /ɒ/

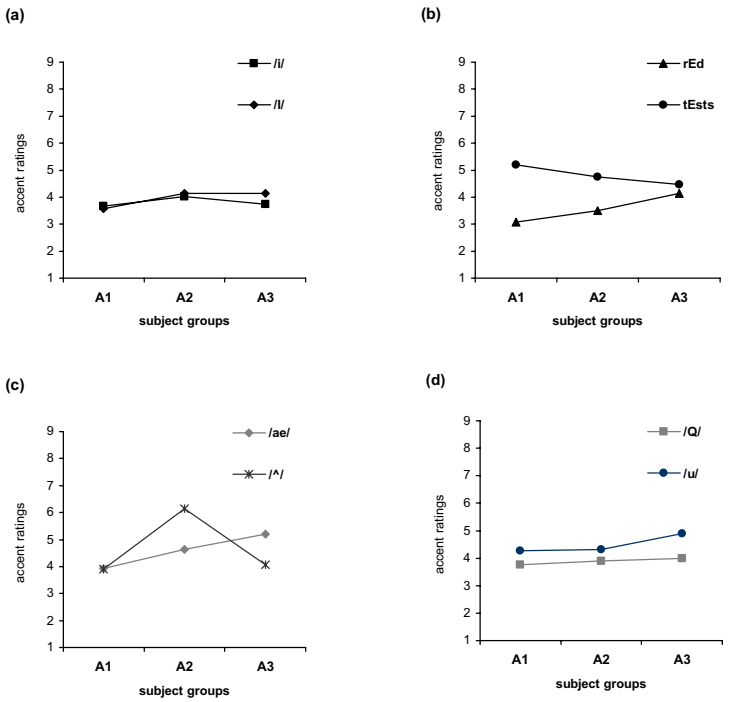
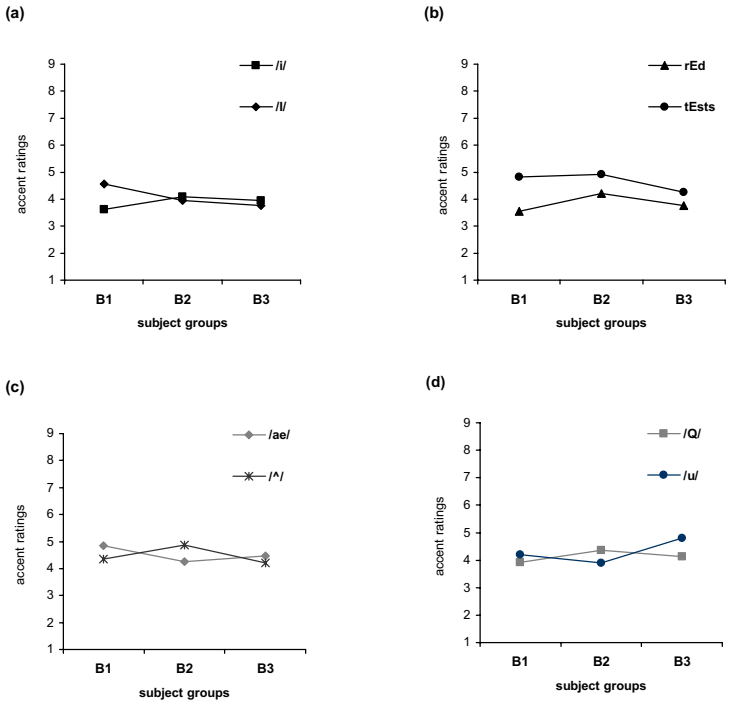


Figure 5.60. FA ratings on group B's production of /i, ɪ, ɛ, æ, ɒ, u, ʌ/. Factor: exposure to the FL. Note: /Q/ = /ɒ/



Figures 5.61 and 5.62 below provide an indication of older learners' production of English vowels, as their experience in the FL increased. Thus, it can be seen that Group C followed the expected pattern, i.e. 14-year-old beginners received less foreign-accented scores on English vowels along with an increase in exposure. So did adult learners in their production of /i/, /ɛ/ (*tests*), /æ/, and /ʌ/, while in the remaining vowel sounds such consistent pattern was not obtained.

Figure 5.61. FA ratings on group C's production of /i, ɪ, ɛ, æ, ɒ, u, ʌ/. Factor: exposure to the FL. Note: /Q/ = /ɒ/

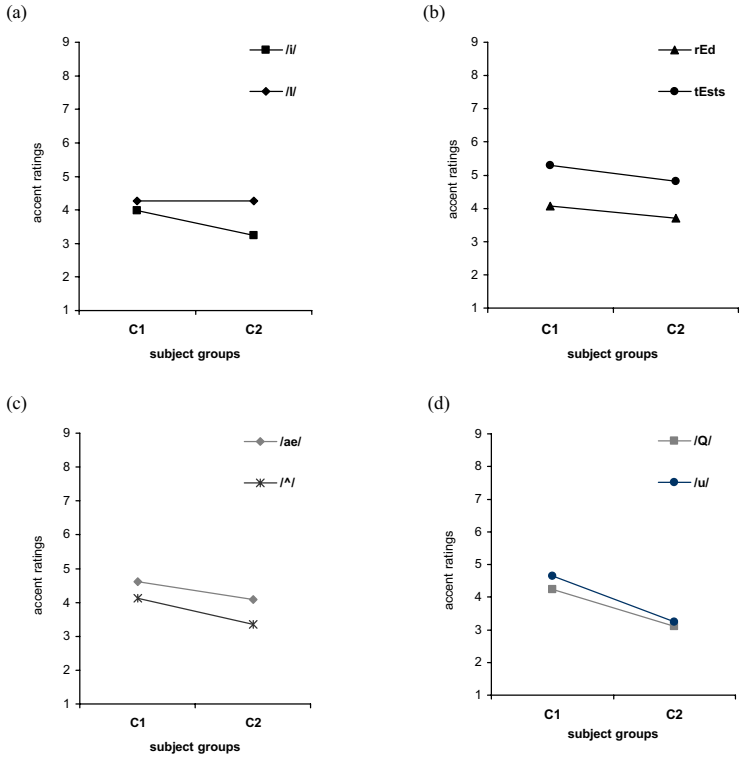


Figure 5.62. FA ratings on group D's production of /i, ɪ, ɛ, æ, ɒ, u, ʌ/. Factor: exposure to the FL. Note: /Q/ = /ɒ/

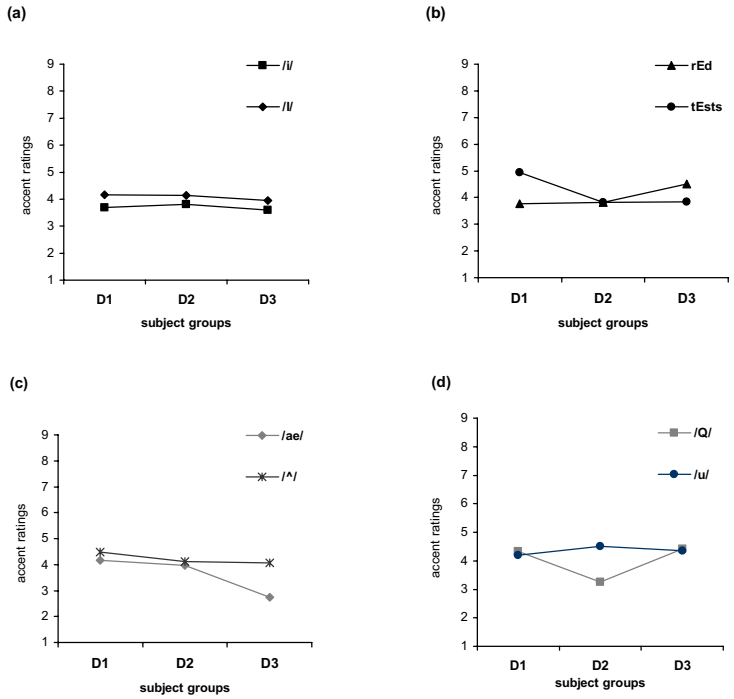


Table 5.49 below summarises the results of the statistical analyses performed on the accent ratings for vowels with exposure to English as a factor.

Table 5.49. Summary of comparisons carried out on the accent ratings on vowels (averaged over judges) with exposure as a factor. Nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each group's mean accent ratings for each vowel appear in parentheses.

(a)

Factor: ***exposure to English***
D.V.: FA ratings on /i/
A1 (3.67) – A2 (4.02) – A3 (3.73) n.s.
B1 (3.62) – B2 (4.10) – B3 (3.95) n.s.

(b)

Factor: ***exposure to English***
D.V.: FA ratings on /ɪ/
A1 (3.57) – A2 (4.13) – A3 (4.14) n.s.
B1 (4.57) – B2 (3.96) – B3 (3.77) n.s.

(c)

Factor: ***exposure to English***
D.V.: FA ratings on /ɛ/ *red*
A1 (3.08) – A2 (3.50) – A3 (4.15) ***p*** = .071
B1 (3.54) – B2 (4.21) – B3 (3.76) n.s.

(d)

Factor: ***exposure to English***
D.V.: FA ratings on /ɛ/ *tests*
A1 (5.21) – A2 (4.75) – A3 (4.47) n.s.
B1 (4.83) – B2 (4.91) – B3 (4.26) n.s.

(e)

Factor: ***exposure to English***
D.V.: FA ratings on /æ/
A1 (3.92) – A2 (4.64) – A3 (5.21) ***p*** = .066
B1 (4.85) – B2 (4.26) – B3 (4.46) n.s.

(f)

Factor: ***exposure to English***
D.V.: FA ratings on /ɒ/
A1 (3.77) – A2 (3.90) – A3 (3.98) n.s.
B1 (3.93) – B2 (4.36) – B3 (4.14) n.s.

(g)

Factor: ***exposure to English***
D.V.: FA ratings on /u/
A1 (4.27) – A2 (4.64) – A3 (4.19) n.s.
B1 (4.21) – B2 (3.89) – B3 (4.80) n.s

(h)

Factor: ***exposure to English***
D.V.: FA ratings on /ʌ/
A1 (3.91) – A2 (4.10) – A3 (4.07) n.s.
B1 (4.34) – B2 (4.86) – B3 (4.20) n.s.

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p < .05$.

5.3.2.2.3. Effect of dominant L1(s)

Both mean accent ratings on each of the vowel sounds averaged over 7 judges and judges' accent ratings averaged over the 7 vowels (in 11 words) were submitted to separate Kruskal-Wallis analyses with dominant L1(s) as a factor (no subject group distinction, except for L1). All the analyses conducted yielded significant differences in the ratings assigned to the several language groups (adjusted $p < .05$). Mann-Whitney *U* tests revealed that English-speaking foils received significantly lower accent scores than Catalan and/or Spanish native speakers in all the dependent variables submitted to the nonparametric statistical analyses. However, among the learners' language groups no difference in ratings was significant – only two comparisons approached significance: the Catalan dominant and Spanish dominant speaker comparison in /æ/ (*U* 411, *Z* –1.707, $p = .088$, adjusted $p > .05$) and that of Spanish dominant speakers and Catalan-Spanish balanced bilinguals in /u/ (*U* 1023.5, *Z* –2.070, $p = .038$, adjusted $p > .05$) (see Tables 5.50 and 5.51). Figures 5.63 and 5.64 display each of the judges' ratings averaged over those vowel segments, in addition to the language subgroups' accent ratings on each of the 7 English segments investigated in this study.

Figure 5.63. Judges' FA ratings averaged over vowels. Factor: dominant L1(s).

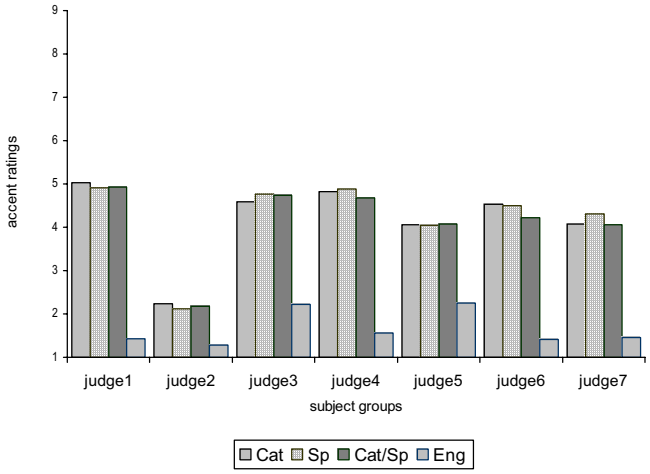


Figure 5.64. FA ratings on /i, ɪ, ɛ, æ, ɒ, u, ʌ/ averaged over judges. Factor: dominant L1(s). Note: /Q/ = /ɒ/

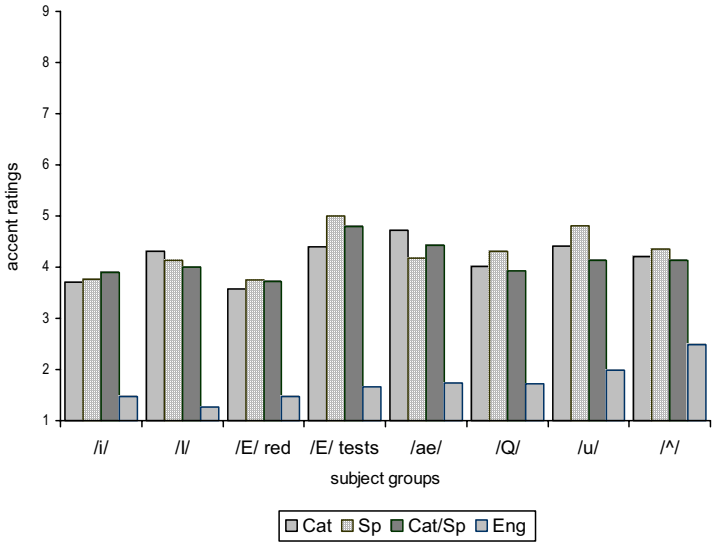


Table 5.50. Summary of comparisons carried out on the judges' accent ratings (averaged over vowels) with dominant L1(s) as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). Each group's mean accent ratings by each judge appear in parentheses.

(a)

Factor: ***dominant L1(s)***
D.V.: Judge 1's FA ratings
Cat (5.03) – Sp (4.91) – C/S (4.92) – NE (1.43)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat – NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(b)

Factor: ***dominant L1(s)***
D.V.: Judge 2's FA ratings
Cat (2.23) – Sp (2.12) – C/S (2.17) – NE (1.28)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat – NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(c)

Factor: ***dominant L1(s)***
D.V.: Judge 3's FA ratings
Cat (4.59) – Sp (4.77) – C/S (4.74) – NE (2.22)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat – NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(d)

Factor: ***dominant L1(s)***
D.V.: Judge 4's FA ratings
Cat (4.82) – Sp (4.88) – C/S (4.68) – NE (1.56)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat – NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(e)

Factor: ***dominant L1(s)***
D.V.: Judge 5's FA ratings
Cat (4.06) – Sp (4.04) – C/S (4.07) – NE (2.25)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat – NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(f)

Factor: ***dominant L1(s)***
D.V.: Judge 6's FA ratings
Cat (4.53) – Sp (4.50) – C/S (4.22) – NE (1.41)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat – NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(g)

Factor: ***dominant L1(s)***
D.V.: Judge 7's FA ratings
Cat (4.07) – Sp (4.31) – C/S (4.06) – NE (1.46)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat – NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p < .05$.

Table 5.51. Summary of comparisons carried out on the accent ratings on vowels (averaged over judges) with dominant L1(s) as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each group's mean accent ratings for each vowel appear in parentheses.

(a)

Factor: ***dominant L1(s)***
D.V.: FA ratings on /i/

Cat (3.70) – Sp (3.77) – C/S (3.90) – NE (1.46)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat – NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(b)

Factor: ***dominant L1(s)***
D.V.: FA ratings on /ɪ/

Cat (4.31) – Sp (4.13) – C/S (4.00) – NE (1.27)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat – NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(c)

Factor: ***dominant L1(s)***
D.V.: FA ratings on /ɛ/ *red*

Cat (3.58) – Sp (3.75) – C/S (3.72) – NE (1.45)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat – NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(d)

Factor: ***dominant L1(s)***
D.V.: FA ratings on /ɛ/ *tests*

Cat (4.39) – Sp (5.00) – C/S (4.79) – NE (1.62)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat – NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(e)

Factor: ***dominant L1(s)***
D.V.: FA ratings on /æ/

Cat (4.72) – Sp (4.18) – C/S (4.43) – NE (1.73)*
Cat – Sp ***p*** = .088
Cat – C/S n.s.
Cat – NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(f)

Factor: ***dominant L1(s)***
D.V.: FA ratings on /ɒ/

Cat (4.02) – Sp (4.31) – C/S (3.93) – NE (1.71)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat – NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(g)

Factor: ***dominant L1(s)***
D.V.: FA ratings on /ʊ/

Cat (4.41) – Sp (4.81) – C/S (4.13) – NE (1.97)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat – NE*
Sp – C/S ***p*** = .038
Sp < NE*
C/S < NE*

(h)

Factor: ***dominant L1(s)***
D.V.: FA ratings on /ʌ/

Cat (4.21) – Sp (4.36) – C/S (4.13) – NE (2.42)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat – NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p < .05$.

The Kruskal-Wallis analyses carried out within each learner group did not yield any significant differences – at the most some results approached significance. Both the results approaching significance and nonsignificant comparisons had a common trait in that all failed to outline a consistent pattern or advantage of a specific language subgroup obtaining higher or lower FA ratings over the remaining language subgroups in their corresponding subject group. Thus, figures and summary tables for each learner group are not presented.

5.3.2.2.4. Effect of gender

Overall, female subjects obtained lower FA ratings than male subjects[98]. The differences in ratings between the two gender groups were significant for judges 1, 3, 6, and 7 (U 12.96.5, Z –3.759, adjusted $p < .05$; U 1120, Z –4.569, adjusted $p < .05$; U 978, Z –5.222, adjusted $p < .05$; U 1249, Z –3.977, adjusted $p < .05$), while they approached significance for the remaining judges (U 1695, Z –1.935, $p = .053$, adjusted $p > .05$ for judge 2; U 1650.5, Z –2.137, $p = .033$, adjusted $p > .05$ for judge 4; and U 1670.5, Z –2.046, $p = .041$, adjusted $p > .05$) (see Figure 5.65 and Table 5.52).

In addition, female ratings obtained for /i/, /ɛ/ (*red*) and /æ/ were significantly lower than male accent ratings (U 1728.5, Z –3.388, adjusted $p < .05$; U 1456, Z –4.552, adjusted $p < .05$; U 1276, Z –4.502, adjusted $p < .05$), and those for /ɛ/ (*tests*) and /ʌ/ approached significance (U 2037.5, Z –2.112, $p = .035$, adjusted $p > .05$; U 1858, Z –2.828, $p = .005$, adjusted $p > .05$) (see Figure 5.66 and Table 5.53).

As a function of learner group, the FA ratings obtained for both male and female speakers agreed with the finding reported above in that female subjects were considered to have produced English vowel segments as less foreign-accented than male subjects. No difference in ratings reached significance, rather they approached significance[99].

[98] As in Study 1, English-speaking Ss were not included in the analyses performed with gender as a factor.
[99] Neither figures nor summary tables are presented with the accent ratings obtained for each learner group on the same grounds as those outlined in Study 1.

Figure 5.65. Judges' FA ratings averaged over vowels. Factor: gender.

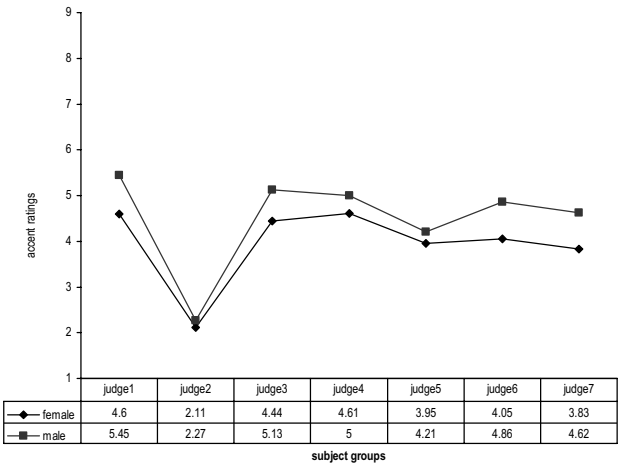


Figure 5.66. FA ratings on /i, ɪ, ɛ, æ, ɒ, u, ʌ/ averaged over judges. Factor: gender. Note: /Q/ = /ɒ/

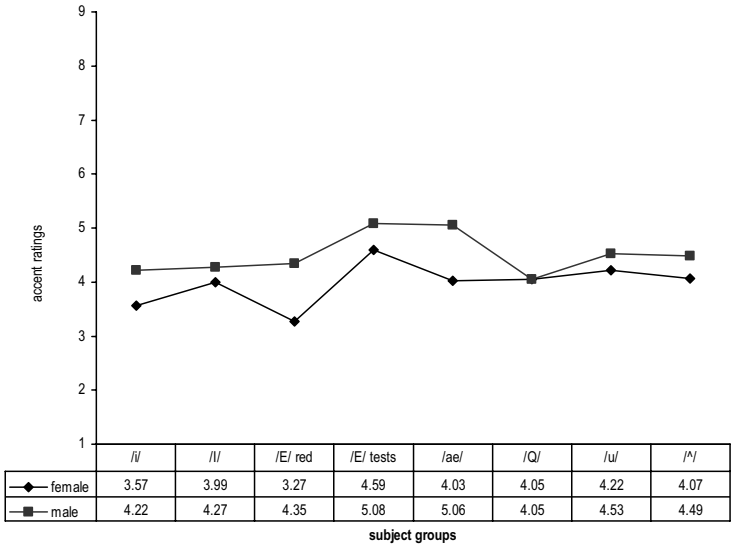


Table 5.52. Summary of comparisons carried out on the judges' accent ratings (averaged over vowels) with gender as a factor. Significant comparisons are marked with * (p < .05), while nonsignificant group comparisons are displayed as n.s. (p > .05). *P* values of marginally significant results (.05 < p < .10) are stated as p = .exact significance value. Each group's mean accent ratings by each judge appear in parentheses.

(a)

Factor: ***gender*** *D.V.:* Judge 1's FA ratings
Female (4.60) > Male (5.45)*

(b)

Factor: ***gender*** *D.V.:* Judge 2's FA ratings
Female (2.11) – Male (2.27) ***p*** = .053

(c)

Factor: ***gender*** *D.V.:* Judge 3's FA ratings
Female (4.44) – Male (5.13)*

(d)

Factor: ***gender*** *D.V.:* Judge 4's FA ratings
Female (4.61) – Male (5.00) ***p*** = .033

(e)

Factor: ***gender*** *D.V.:* Judge 5's FA ratings
Female (3.95) – Male (4.21) ***p*** = .041

(f)

Factor: ***gender*** *D.V.:* Judge 6's FA ratings
Female (4.05) – Male (4.86)*

(g)

*Factor:****gender*** *D.V.:* Judge 7's FA ratings
Female (3.83) – Male (4.62)*

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted p < .05. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted p < .05.

Table 5.53. Summary of comparisons carried out on the accent ratings on vowels (averaged over judges) with gender as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each group's mean accent ratings for each vowel appear in parentheses.

(a)

Factor: ***gender***
D.V.: FA ratings on /i/
Female (3.57) – Male (4.22)*

(b)

Factor: ***gender***
D.V.: FA ratings on /ɪ/
Female (3.99) – Male (4.27) n.s.

(c)

Factor: ***gender***
D.V.: FA ratings on /ɛ/ *red*
Female (3.27) – Male (4.35)*

(d)

Factor: ***gender***
D.V.: FA ratings on /ɛ/ *tests*
Female (4.59) – Male (5.08) ***p*** = .033

(e)

Factor: ***gender***
D.V.: FA ratings on /æ/
Female (4.03) – Male (5.06)*

(f)

Factor: ***gender***
D.V.: FA ratings on /ɒ/
Female (4.05) – Male (4.05) n.s.

(g)

Factor: ***gender***
D.V.: FA ratings on /u/
Female (4.22) – Male (4.53) n.s.

(h)

Factor: ***gender***
D.V.: FA ratings on /ʌ/
Female (4.07) – Male (4.49) ***p*** = .041

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .001. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .001 as the resulting adjusted $p < .05$.

5.3.2.3. Vowel identification task /i, ɪ, ɛ, æ, ɒ, u, ʌ/

A total of 14,357 vowel identifications was obtained, which included both correct identification and misidentification scores. As mentioned above (5.3.1.6. Procedure), correct identification scores referred to those instances where listeners identified the target sound as intended. That is, the sound in question was identified as either a good instance, slightly distorted, or very distorted realisation of the specific segment being examined. Misidentification scores, on the other hand, involved those vowel productions (or substitutions) that listeners did not characterise as intended. Each of the scores was based on 7 judgements (7 listeners x 1 presentation of each subject's vowel production). Next, identification and misidentification scores were converted to percent scores, so that comparisons could be made across the various subject groups.

Figures 5.67 through 5.77 below display each subject group's percent correct identification scores for each vowel segment examined (in each of the 11 words), as well as the subjects' substitutions (or mispronunciations) for the target sounds as identified by the seven judges. A first look at these figures allows us to see that the control group generally obtained higher frequency correct scores than learners of English[100] (range = 73.62% – 98.90% for English foils[101] vs. 38.46% – 91.59% for FL learners).

As for substitutions, the following general trends can be observed. First, in the case of /i/, [i] (non-diphthongised) and [ɪ] were found to be the most frequent substitutions, whereas for /ɪ/ both [i] and [ij] (diphthongised) were heard more frequently instead of the target sound. Like /i/, the most frequent substitute for /u/ was the corresponding non-diphthongised sound [u] and, to a lesser extent, [ʊw] and [ɔ]. [ɪ] and [e] were the most frequent substitutes for /ɛ/, while [a] and [ɑ] often substituted for the target /æ/. As for /ɒ/, more substitutions were heard than those reported above for the

[100] As in the FA rating task, the identification scores that C2, D2, and D3 obtained are only included as an indication of those learners' performance. But due to the small number of subjects comprising those groups, their results are not analysed statistically.

[101] Somewhat surprising, though, was the extent to which the control group's /ɒ/s and /ʌ/s were identified as intended (mean percent: 73.62 and 74.72, respectively). In spite of this, the identification scores obtained in their production of /ɒ/ and /ʌ/ were higher than those of learners (see Figures 5.75 and 5.77).

other segments with a frequency rate higher than 2%[102]: [ɑ], [ɔ], [ʌ], [ʊ], and [u]. Likewise, /ʌ/ presented a variety of misidentification patterns: [æ], [a], [ɒ], [ɔ], and [ʊ].

As a function of dominant L1(s), NE obtained higher correct identification scores in the production of all vowels than learners of English. Only in the production of /ɛ/ did learners' scores closely resemble those of English foils. Among the three language subgroups – Catalan and Spanish dominant speakers, and Catalan/Spanish balanced bilinguals – there was no clear advantage of one language group over the remaining groups in the production of English vowels. By contrast, the misidentification patterns for each of the 7 vowels were identical across the three language groups. Moreover, the most frequent substitutions coincided with those reported above.

When looking at the variable of gender, the higher percent correct identification scores obtained by female speakers agree with the findings of FA ratings on vowels reported in the first part of Study 2. Both female and male subjects obtained similar rates of the most frequent substitutions – which, in turn, agreed with the misidentification patterns so far described.

[102] Figures 5.67 – 5.77 show all the results (identifications and misidentifications) obtained. However, only sounds with a frequency of appearance above 2% were considered in the statistical analyses conducted and subsequent figures, following previous studies (e.g. Flege, 1991a).

Figure 5.67. Correct identification scores and misidentifications for /i/ (*speak*). Results for C2, D2, and D3 are indicative only.

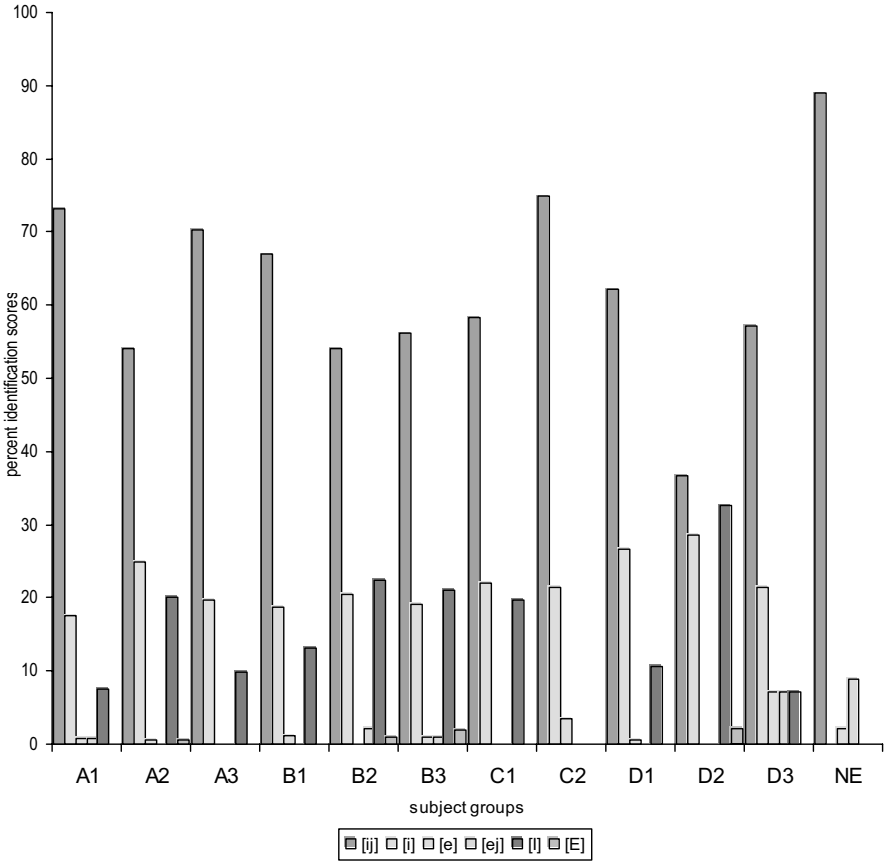


Figure 5.68. Correct identification scores and misidentifications for /i/ (*tea*). Results for C2, D2, and D3 are indicative only.

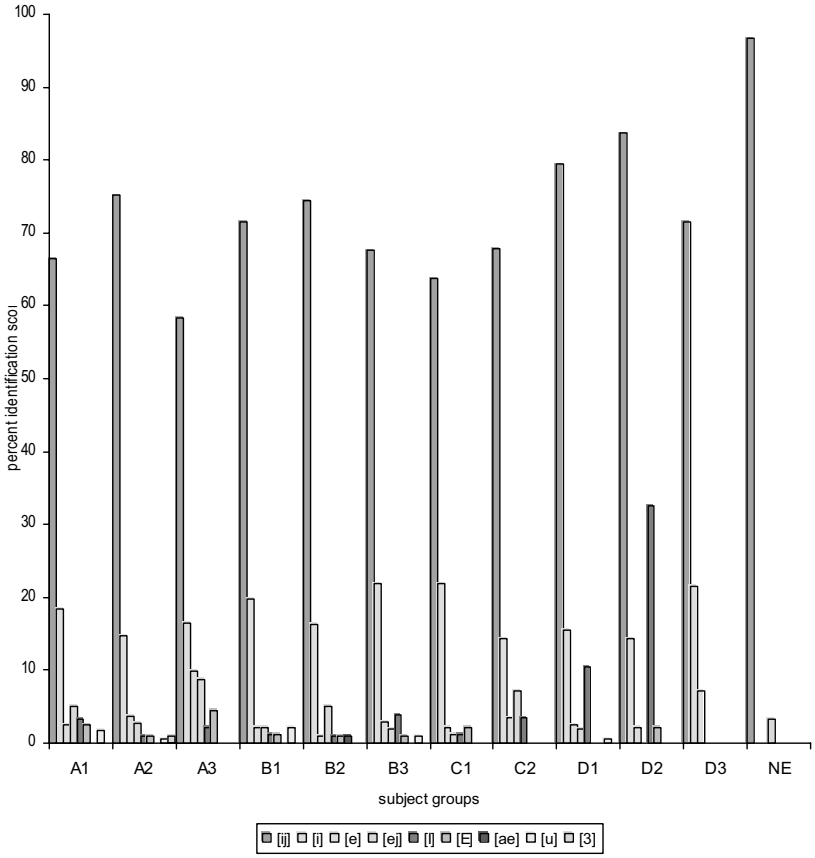


Figure 5.69. Correct identification scores and misidentifications for /ɪ/ (*it*). Results for C2, D2, and D3 are indicative only.

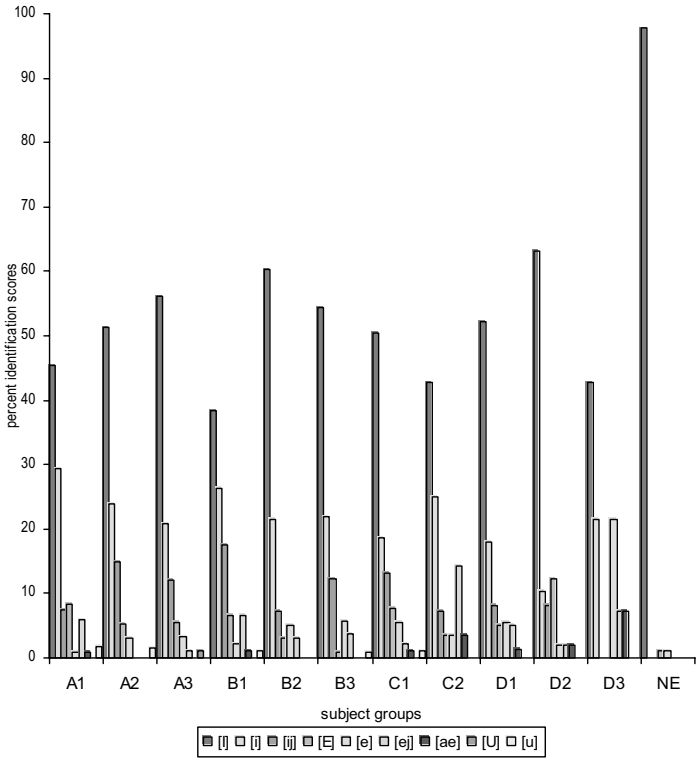


Figure 5.70. Correct identification scores and misidentifications for /ɪ/ (*this*). Results for C2, D2, and D3 are indicative only.

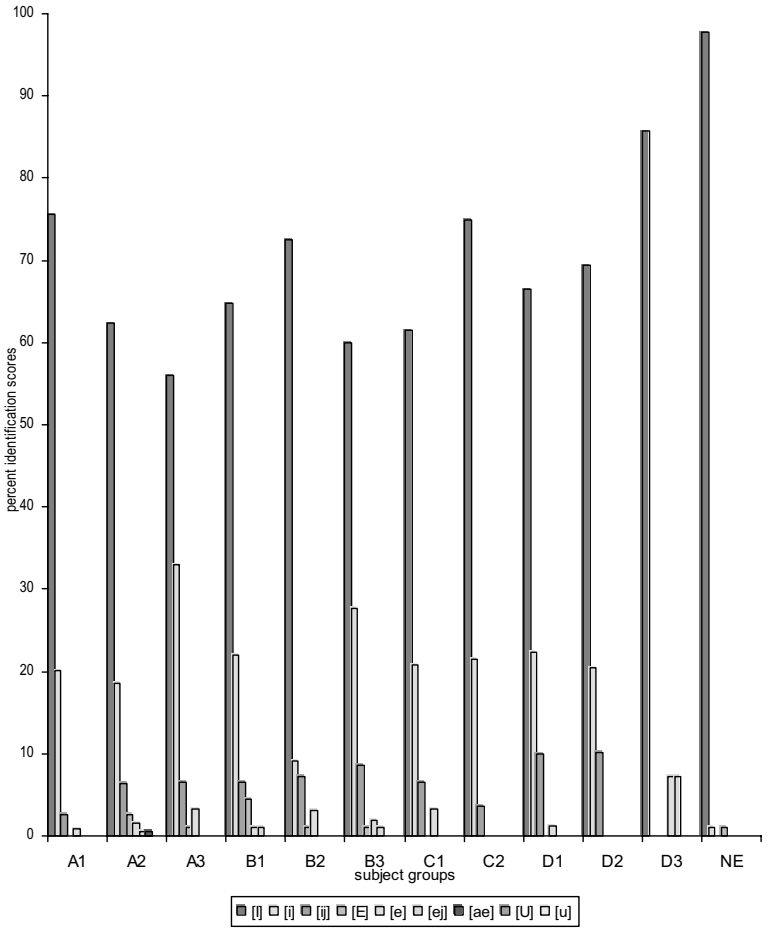


Figure 5.71. Correct identification scores and misidentifications for /ɛ/ (*red*). Results for C2, D2, and D3 are indicative only.

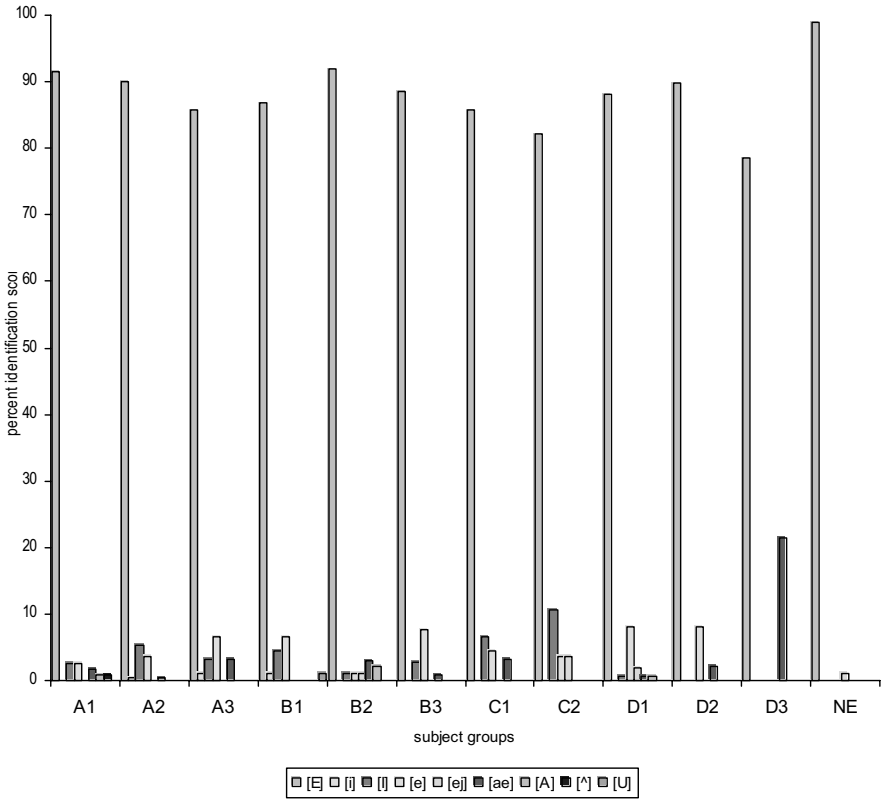


Figure 5.72. Correct identification scores and misidentifications for /ɛ/ (*tests*). Results for C2, D2, and D3 are indicative only.

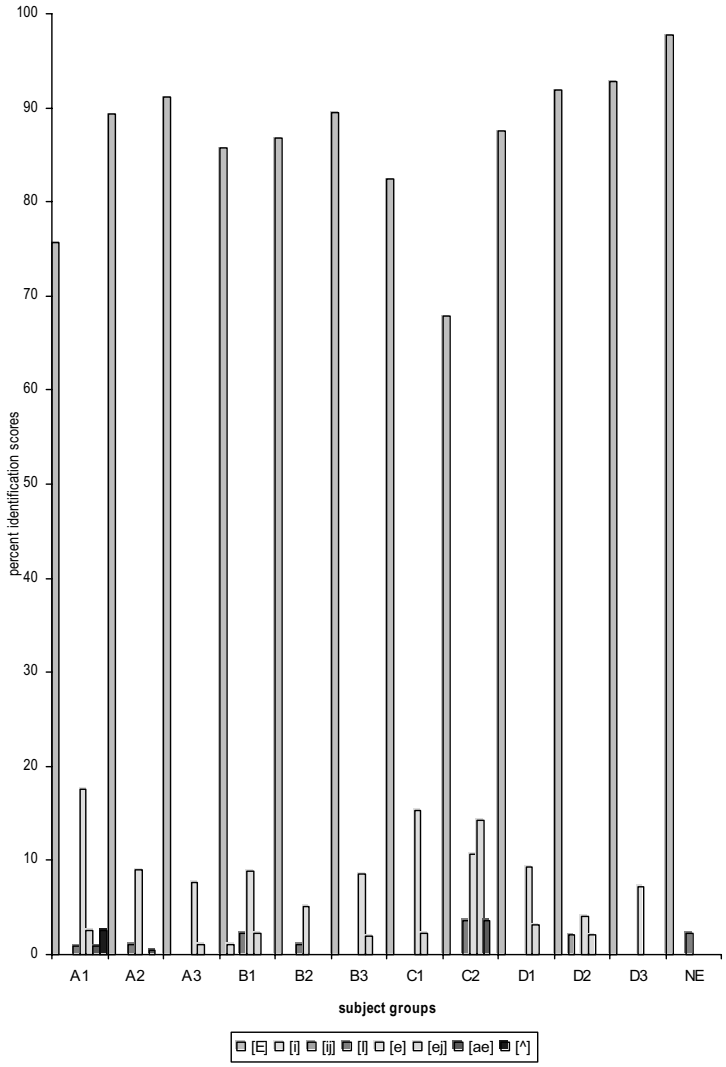


Figure 5.73. Correct identification scores and misidentifications for /æ/ (*back*). Results for C2, D2, and D3 are indicative only.

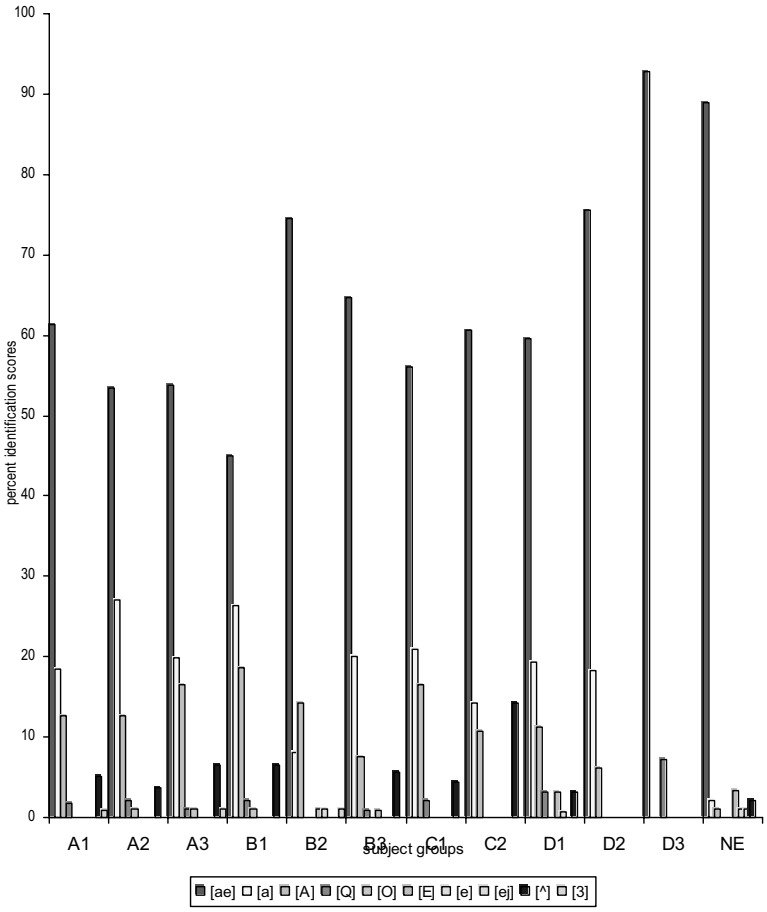


Figure 5.74. Correct identification scores and misidentifications for /æ/ (*pad*). Results for C2, D2, and D3 are indicative only.

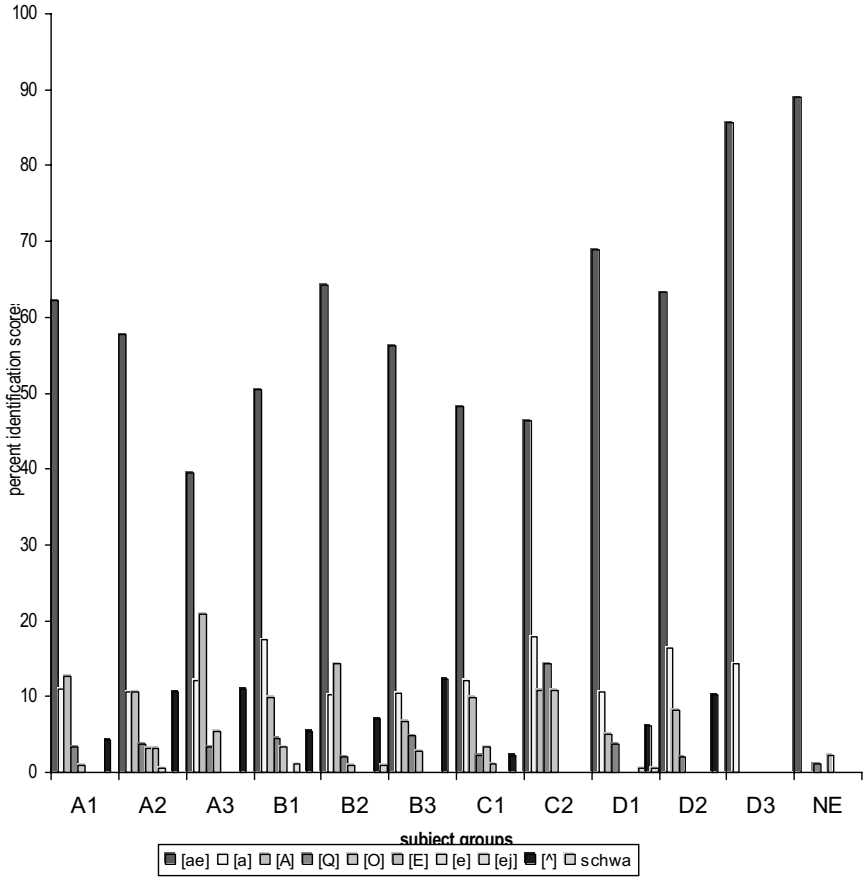


Figure 5.75. Correct identification scores and misidentifications for /ɒ/ (*box*). Results for C2, D2, and D3 are indicative only.

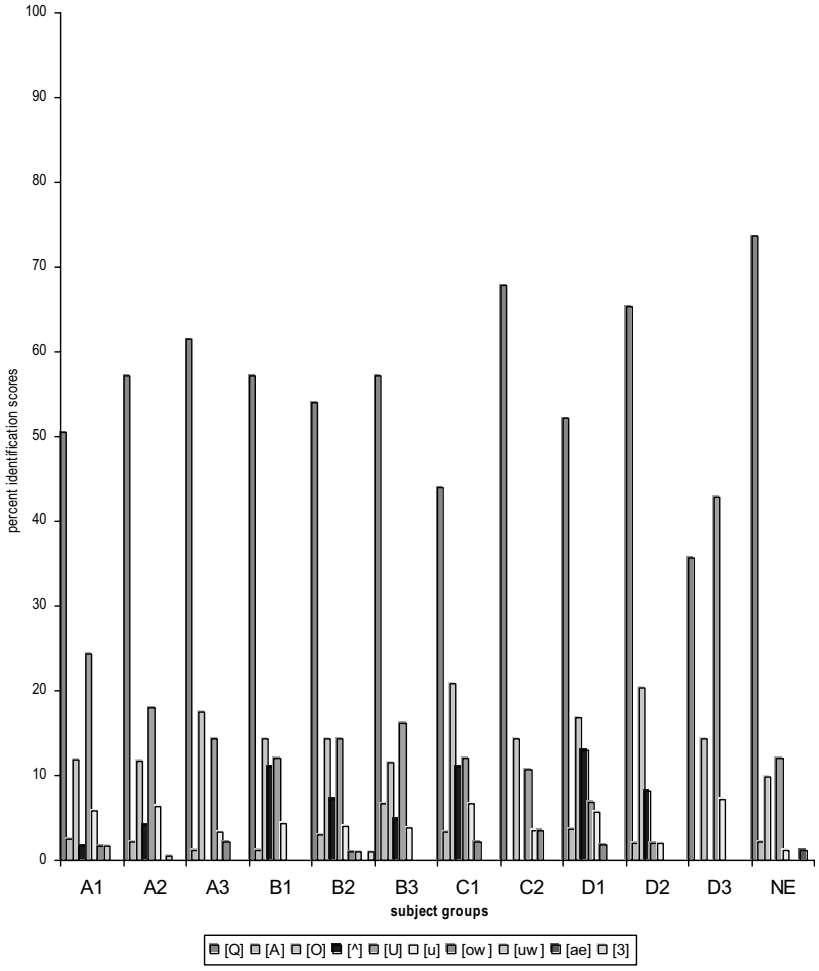


Figure 5.76. Correct identification scores and misidentifications for /u/ (*zoo*). Results for C2, D2, and D3 are indicative only.

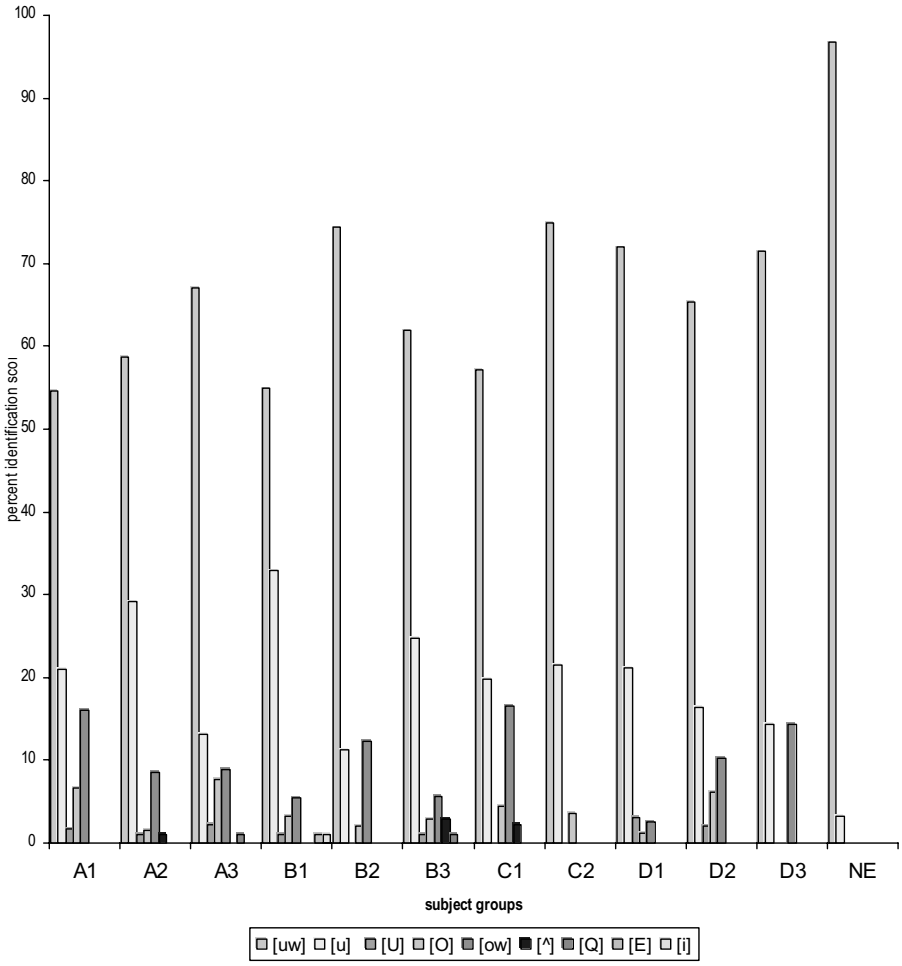
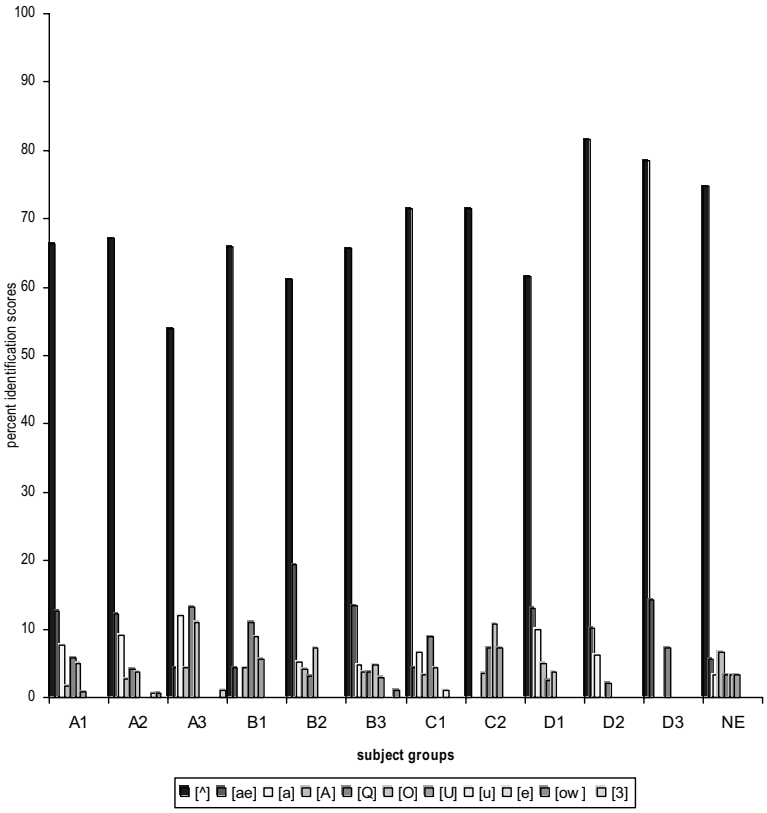


Figure 5.77. Correct identification scores and misidentifications for /ʌ/ (*but*). Results for C2, D2, and D3 are indicative only.



Averaged over all participants, the percent correct identifications for each vowel segments were as follows: 62.73 and 73.38 for /i/ (in *speak* and *tea*, respectively), 55.10 and 68.41 for /ɪ/ (in *it* and *this*, respectively), 89.35 and 87.04 for /ɛ/ (in *red* and *tests*, respectively), 62.11 and 60.07 for /æ/ (in *back* and *pad*, respectively), 56.43 for /ɒ/ (in *box*), 66.01 for /u/ (in *zoo*), and 66.28 for /ʌ/ (in *but*).

Those percent scores on each vowel sound were submitted to a repeated measures ANOVA with age of FL learning (5 levels) x exposure (4 levels) as factors and identification scores as repeated measures. The simple main effect of scores was significant (F (10, 140) = 23.786, $p < .001$). No other simple effect or interaction was significant. Pairwise vowel comparisons showed that /ɪ/ in *it* received significantly lower correct identification scores than /i/, /ɪ/ (*this*), /ɛ/, and /ʌ/ ($p < .05$). On the contrary, /ɛ/ was identified as intended at significantly higher rates than the other six vowels. Homogeneity tests further showed that correct identification scores were not always normally distributed, as was the case of /i/ (*tea*), /ɪ/ (*it*, *this*), and /u/ (*zoo*) ($p < .05$).

Based on the significant differences found in scores, as well as Levene's Tests of Equality of Error variances, the subsequent analyses performed were nonparametric. According to Wilcoxon Signed Ranks Tests, correct identification scores for /ɛ/ in *red* and *tests*, on the one hand, and for /æ/ in *back* and *pad*, on the other hand, were not significantly different (Z –1.219, $p = .223$; and Z –.309, $p = .757$, respectively); whereas for /i/ in *speak* and *tea*, and /ɪ/ in *it* and *this*, the differences in scores in the words each sound appeared were significant (Z –3.510, $p < .001$; and Z –4.516, $p < .001$, respectively). Therefore, data could only be grouped together in /ɛ/ and /æ/ when determining the effects of the variables in the study. Furthermore, to study the differences (or lack of differences) in scores among three or more groups, Kruskal-Wallis analyses were conducted on the identification scores and with age of FL learning, exposure, dominant L1(s), and gender as factors (one factor at a time). In the event of significant results, Mann-Whitney U tests were carried out in order to find out where exactly the differences occurred.

Last, in order to conduct all these analyses, a significance level of .007 was adopted, so that the experiment-wise error was held constant at .05 (.007 x 7 vowel segments).

5.3.2.3.1. Effect of onset age of FL learning

As shown in Figures 5.78 – 5.86, the control group of English-speaking subjects obtained correct identification scores at higher frequency rates than learner groups. Kruskal-Wallis analyses conducted on the correct identification scores for each vowel sound as dependent variables and onset age of FL learning as a factor showed that differences in scores between the NE group and learner groups were for the most part significant (adjusted $p < .05$). More precisely, the differences in scores for /i/ (*tea*), /ɪ/, /æ/, and /u/ were always significant. Moreover, scores for /i/ (*speak*) and /ɛ/ were nearly always significant – otherwise, they approached significance, as was the case of the A1–B1–C1–D1–NE and A3–B3–NE comparisons in /i/ (*speak*) (χ^2 10.364, *df* 4, $p = .035$, adjusted $p > .05$; and χ^2 7.563, *df* 2, $p = .023$, adjusted $p > .05$, respectively), and that of A2–B2–NE in /ɛ/ (χ^2 8.781, *df* 2, $p = .012$, adjusted $p > .05$). Mann-Whitney *U* tests further indicated that the differences in scores for /u/ between B2 and NE approached significance (*U* 56, *Z* –1.964, $p = .050$, adjusted $p > .05$), as well as those for /ɪ/ (*this*) between A1 and NE, on the one hand, and between B1 and NE, on the other (*U* 63, *Z* –2.309, $p = .021$, adjusted $p > .05$; and *U* 46, *Z* –2.328, $p = .020$, adjusted $p > .05$). As for /ɒ/, results only approached significance in the A1–B1–C1–D1–NE (χ^2 9.551, *df* 4, $p = .049$, adjusted $p > .05$). The scores obtained for /ʌ/ were pretty similar across all subject groups (including NE foils) resulting in the lack of significant differences among them (adjusted $p > .05$).

Among learner groups, Mann-Whitney *U* tests showed that identification scores obtained by A1, B1, C1, and D1 were not significantly different (adjusted $p > .05$), although a few comparisons approached significance: A1–D1 in /u/ (*U* 128.5, *Z* –1.868, $p = .062$), B1–D1 in /æ/ and /u/ (*U* 91.5, *Z* –1.929, $p = .054$; and *U* 91.5, *Z* –1.939, $p = .053$, respectively); and C1–D1 in /i/ (*tea*) (*U* 95, *Z* –1.877, $p = .060$). In all cases, adult

learners' (D1) vowels were identified as intended at higher frequency rates than the remaining age groups.

When learners had 416 hours of experience in the TL, the A2–B2 comparison only approached significance in the scores for /æ/ (U 116.5, Z –2.010, p = .044, adjusted p > .05) and /u/ (U 126, Z –2.010, p = .078, adjusted p > .05) in favour of older child starters (B2).

Finally, A3 and B3 (8-year-old and 11-year-old starters with 726 hours of instruction in English, respectively) produced English vowels as intended with very similar correct identification scores, as no difference in scores between A3 and B3 was significant, according to Mann-Whitney U tests (adjusted p > .05) (see also Table 5.54).

Figure 5.78. Correct identification scores for /i/ (*speak*). Factor: onset age of FL learning. Results for C2, D2, and D3 are indicative only.

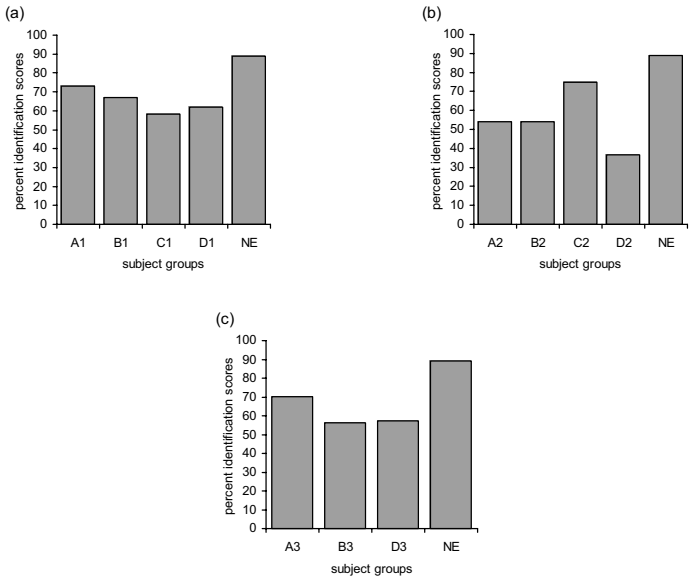


Figure 5.79. Correct identification scores for /i/ (*tea*). Factor: onset age of FL learning. Results for C2, D2, and D3 are indicative only.

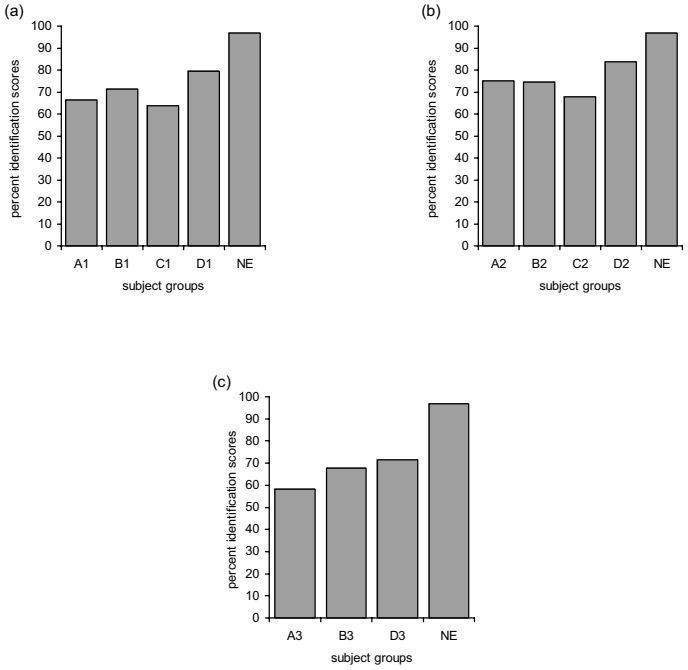


Figure 5.80. Correct identification scores for /ɪ/ (*it*). Factor: onset age of FL learning. Results for C2, D2, and D3 are indicative only.

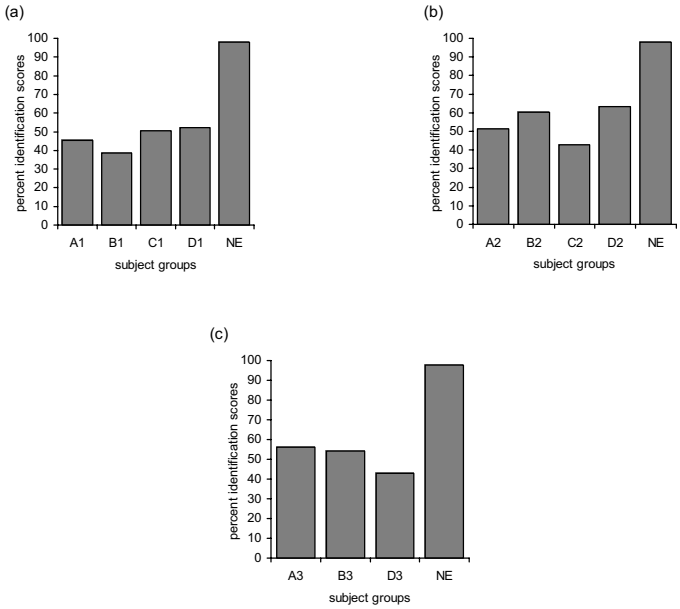


Figure 5.81. Correct identification scores for /ɪ/ (*this*). Factor: onset age of FL learning. Results for C2, D2, and D3 are indicative only.

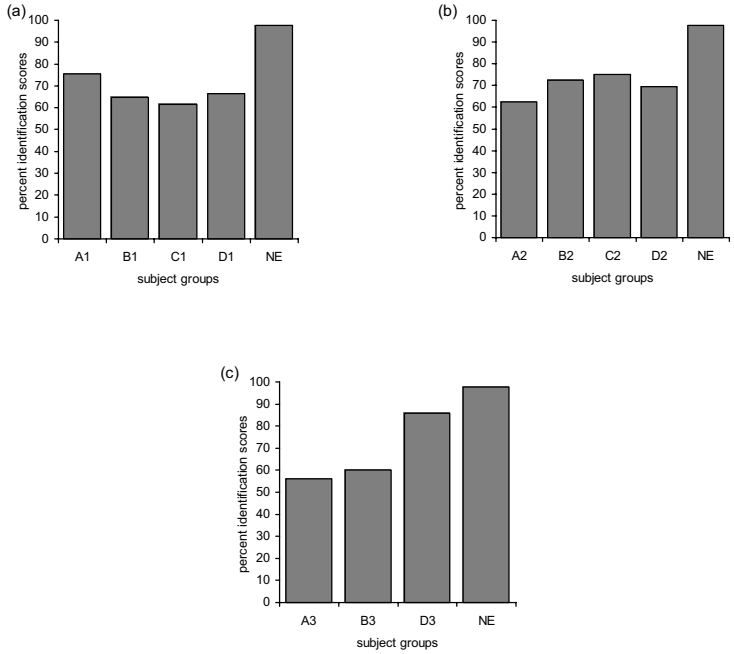


Figure 5.82. Correct identification scores for /ɛ/ (*red* and *tests*). Factor: onset age of FL learning. Results for C2, D2, and D3 are indicative only.

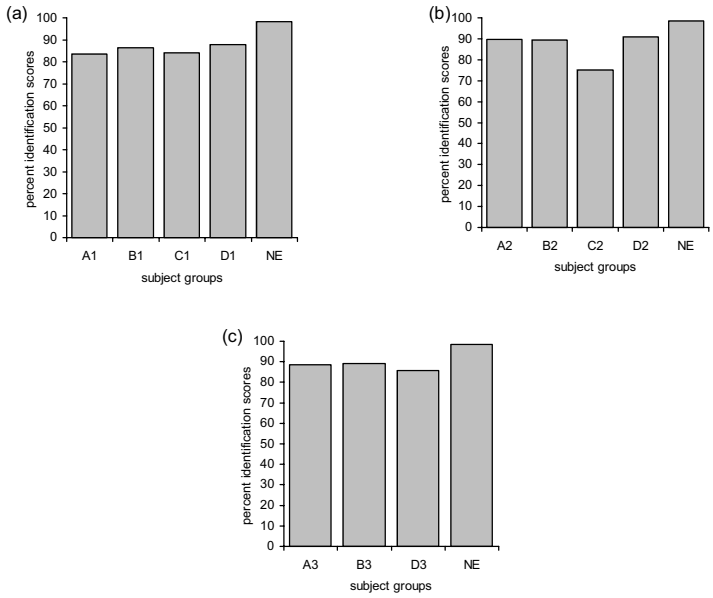


Figure 5.83. Correct identification scores for /æ/ (*back* and *pad*). Factor: onset age of FL learning. Results for C2, D2, and D3 are indicative only.

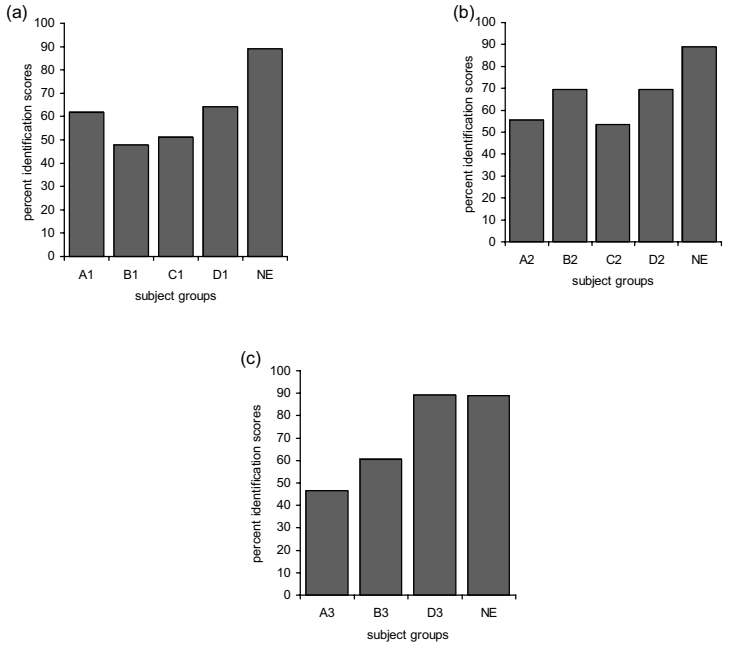


Figure 5.84. Correct identification scores for /ɒ/ (*box*). Factor: onset of FL learning. Results for C2, D2, and D3 are indicative only.

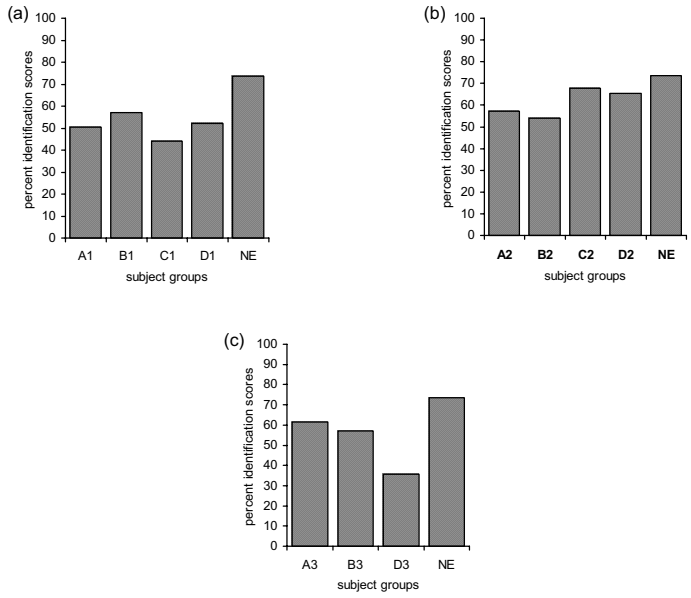


Figure 5.85. Correct identification scores for /u/ (*zoo*). Factor: onset age of FL learning. Results for C2, D2, and D3 are indicative only.

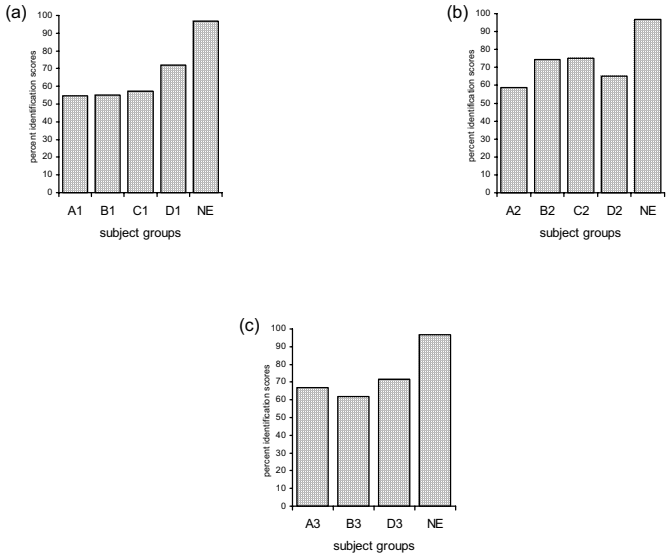


Figure 5.86. Correct identification scores for /ʌ/ (*but*). Factor: onset age of FL learning. Results for C2, D2, and D3 are indicative only.

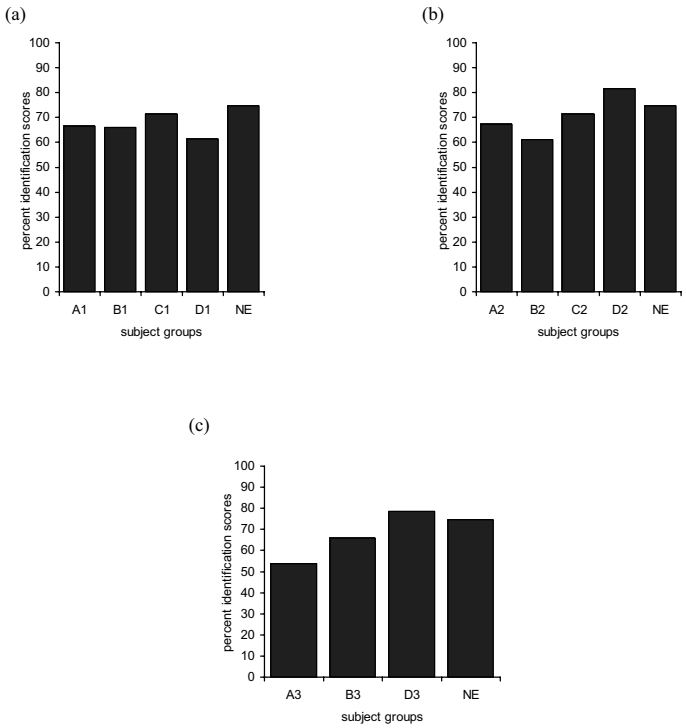


Table 5.54. Summary of comparisons carried out on the vowel percent correct identifications with AOL as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each group's mean percent correct identification scores for each vowel appear in parentheses.

(a)

Factor: ***onset age of FL learning***
D.V.: /i/ (*speak*) percent correct identifications

A1 (73.1) – B1 (67.03) – C1 (58.24) – D1 (62.11) – NE (89.01) ***p*** = .035

A2 (53.96) – B2 (54.08) – NE (89.01)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

A3 (70.32) – B3 (56.19) – NE (89.01) ***p*** = .023

(b)

Factor: ***onset age of FL learning***
D.V.: /i/ (*tea*) percent correct identifications

A1 (66.38) – B1 (71.42) – C1 (63.73) – D1 (79.50) – NE (96.70)*
A1 – B1 n.s.
A1 – C1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – C1 n.s.
B1 – D1 n.s.
B1 < NE*
C1 < D1 ***p*** = .060
C1 < NE*
D1 < NE*

A2 (75.13) – B2 (74.48) – NE (96.70)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

A3 (58.24) – B3 (67.61) – NE (96.70)*
A3 – B3 n.s.
A3 < NE*
B3 < NE*

(c)

Factor: ***onset age of FL learning***
D.V.: /ɪ/ (*it*) percent correct identifications

A1 (45.37) – B1 (38.46) – C1 (50.54) – D1 (52.17) – NE (97.80)*
A1 – B1 n.s.
A1 – C1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – C1 n.s.
B1 – D1 n.s.
B1 < NE*
C1 – D1 n.s.
C1 < NE*
D1 < NE*

A2 (51.32) – B2 (60.20) – NE (97.80)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

A3 (56.04) – B3 (54.28) – NE (97.80)*
A3 – B3 n.s.
A3 < NE*
B3 < NE*

(d)

Factor: ***onset age of FL learning***
D.V.: /ɪ/ (*this*) percent correct identifications

A1 (75.63) – B1 (64.83) – C1 (61.53) – D1 (66.45) – NE (97.80)*
A1 – B1 n.s.
A1 – C1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – C1 n.s.
B1 – D1 n.s.
B1 < NE*
C1 – D1 n.s.
C1 < NE*
D1 < NE*

A2 (62.43) – B2 (72.44) – NE (97.80)*
A2 – B2 n.s.
A2 < NE*
B2 < NE*

A3 (56.04) – B3 (60.00) – NE (97.80)*
A3 – B3 n.s.
A3 < NE*
B3 < NE*

Table 5.54 (continued)

(e)

Factor: ***onset age of FL learning***
D.V.: /ɛ/ percent correct identifications

A1 (83.61) – B1 (86.26) – C1 (84.06) – D1 (87.88) – NE (98.35)*
A1 – B1 n.s.
A1 – C1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – C1 n.s.
B1 – D1 n.s.
B1 < NE*
C1 – D1 n.s.
C1 < NE*
D1 < NE*

A2 (89.67) – B2 (89.28) – NE (98.35) ***p*** = .012

A3 (88.45) – B3 (89.04) – NE (98.35)*
A3 – B3 n.s.
A3 < NE*
B3 < NE*

(f)

Factor: ***onset age of FL learning***
D.V.: /æ/ percent correct identifications

A1 (61.76) – B1 (47.79) – C1 (52.09) – D1 (64.28) – NE (89.01)*
A1 – B1 n.s.
A1 – C1 n.s.
A1 – D1 n.s.
A1 < NE*
B1 – C1 n.s.
B1 < D1 ***p*** = .054
B1 < NE*
C1 – D1 n.s.
C1 < NE*
D1 < NE*

A2 (55.55) – B2 (69.38) – NE (89.01)*
A2 – B2 ***p*** = .044
A2 < NE*
B2 < NE*

A3 (46.70) – B3 (60.47) – NE (89.01)*
A3 – B3 n.s.
A3 < NE*
B3 < NE*

(g)

Factor: ***onset age of FL learning***
D.V.: /ɒ/ percent correct identifications

A1 (50.42) – B1 (57.14) – C1 (43.95) – D1 (52.17) – NE (73.62) ***p*** = .049

A2 (57.14) – B2 (54.08) – NE (73.62) n.s.

A3 (61.53) – B3 (57.14) – NE (73.62) n.s.

(h)

Factor: ***onset age of FL learning***
D.V.: /u/ percent correct identifications

A1 (54.62) – B1 (54.94) – C1 (57.14) – D1 (72.04) – NE (96.70)*
A1 – B1 n.s.
A1 – C1 n.s.
A1 < D1 ***p*** = .062
A1 < NE*
B1 – C1 n.s.
B1 < D1 ***p*** = .053
B1 < NE*
C1 – D1 n.s.
C1 < NE*
D1 < NE*

A2 (58.73) – B2 (74.48) – NE (96.70)*
A2 – B2 p 0 .078
A2 < NE*
B2 < NE ***p*** = .050

A3 (67.03) – B3 (61.90) – NE (96.70)*
A3 – B3 n.s.
A3 < NE*
B3 < NE*

(i)

Factor: ***onset age of FL learning***
D.V.: /ʌ/ percent correct identifications

A1 (66.38) – B1 (65.93) – C1 (71.42) – D1 (61.49) – NE (74.72) n.s.

A2 (67.19) – B2 (61.22) – NE (74.72) n.s.

A3 (53.84) – B3 (65.71) – NE (74.72) n.s.

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .007. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .007 as the resulting adjusted $p < .05$.

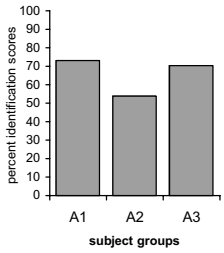
5.3.2.3.2. Effect of exposure

Looking at Figures 5.87 and 5.88, it can be observed that increased hours of instruction in the FL that the subjects had received did not confer an advantage in the percent correct identification scores for each vowel sound examined. Moreover, within each age group, none of the differences in scores between groups reached significance (see Table 5.55 below).

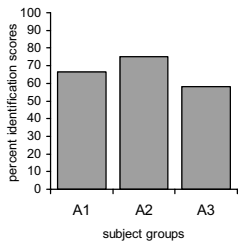
Only in 8-year-old beginners with varying degrees of exposure to English, scores for /i/ (*speak*) and /ɛ/ approached significance (χ^2 5.465, *df* 2, $p = .065$, adjusted $p > .05$; χ^2 6.321, *df* 2, $p = .042$, adjusted $p > .05$). In these two cases, the direction of the effect of exposure was opposite or contradictory: on the one hand, A2's identification scores for /ɛ/ were higher than those of A1 (M = 89.68 vs. 83.61); and, on the other hand, A2 obtained lower percent identification scores for /i/ (53.96) than A1 and A3 (73.1 and 70.32, respectively) (see Figure 5.87).

Figure 5.87. Group A's correct identification scores for /i, ɪ, ɛ, æ, ɒ, u, ʌ/. Factor: exposure to the FL.

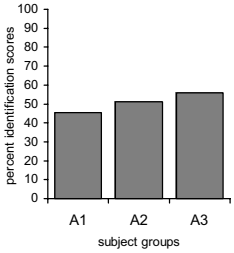
(a)



(b)



(c)



(d)

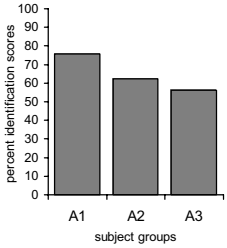
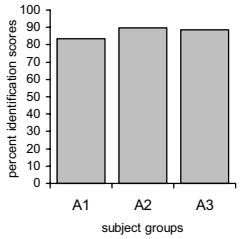
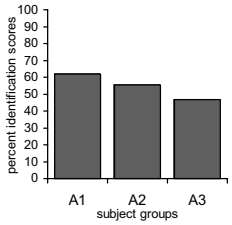


Figure 5.87 (continued)

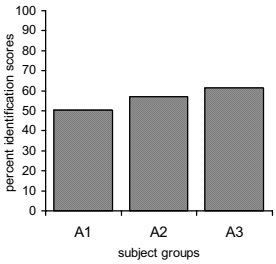
(e)



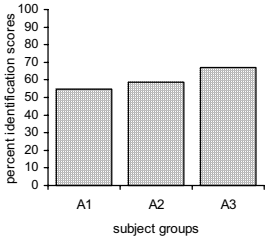
(f)



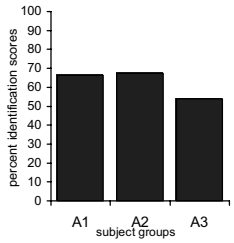
(g)



(h)



(i)



There were no significant differences in the scores obtained by 11-year-old beginners with varying degrees of exposure to the FL, according to Kruskal-Wallis analyses (adjusted $p > .05$) (see Figure 5.88 below).

As noted earlier, the performance of groups C2, D2, and D3 was not analysed statistically due to the small number of subjects. At a descriptive level, results were inconclusive, as well. Therefore, an increase in the number of instruction hours led to higher vowel identification scores, whereas in some other instances the reverse applied, all depending on vowel segment. Thus, 14-year-old beginners and adult starters obtained the same pattern of inconclusive exposure effects as younger beginners.

Figure 5.88. Group B's correct identification scores for /i, ɪ, ɛ, æ, ɒ, u, ʌ/. Factor: exposure to the FL.

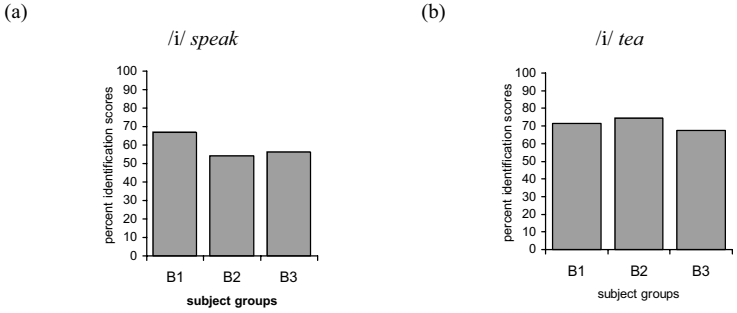
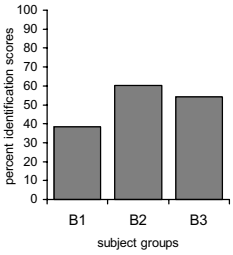


Figure 5.88 (continued)

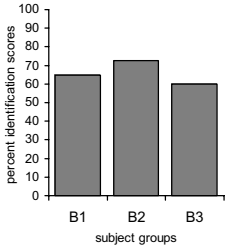
(c)

/ɪ/ it



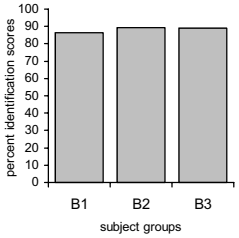
(d)

/ɪ/ *this*



(e)

/ɛ/ *red*, *tests*



(f)

/æ/ *back pad*

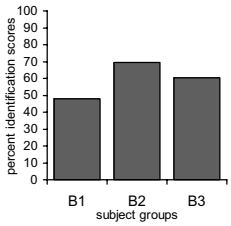
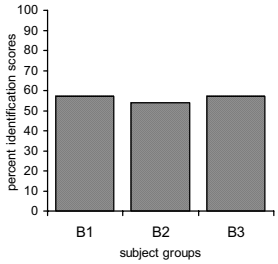


Figure 5.88 (continued)

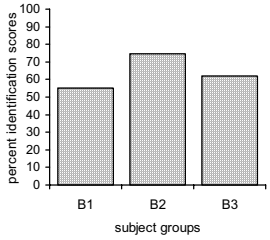
(g)

/ɒ/ *box*



(h)

/u/ *zoo*



(i)

/ʌ/ *but*

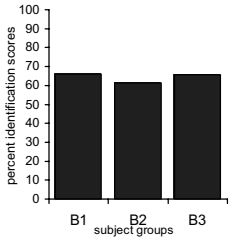


Table 5.55. Summary of comparisons carried out on the vowel percent correct identifications with exposure as a factor. Nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as $p =$ .exact significance value. Each group's mean percent correct identification scores for each vowel appear in parentheses.

(a)

Factor: ***exposure to English***
D.V.: /i/ (*speak*) percent correct identifications
A1 (73.10) – A2 (53.96) – A3 (70.32) ***p*** = .065
B1 (67.03) – B2 (54.08) – B3 (56.19) n.s.

(b)

Factor: ***exposure to English***
D.V.: /i/ (*tea*) percent correct identifications
A1 (66.38) – A2 (75.13) – A3 (58.24) n.s.
B1 (71.42) – B2 (74.48) – B3 (67.61) n.s.

(c)

Factor: ***exposure to English***
D.V.: /ɪ/ (*it*) percent correct identifications
A1 (45.37) – A2 (51.32) – A3 (56.04) n.s.
B1 (38.46) – B2 (60.20) – B3 (54.28) n.s.

(d)

Factor: ***exposure to English***
D.V.: /ɪ/ (*this*) percent correct identifications
A1 (75.63) – A2 (62.43) – A3 (56.04) n.s.
B1 (64.83) – B2 (72.44) – B3 (60.00) n.s.

(e)

Factor: ***exposure to English***
D.V.: /ɛ/ percent correct identifications
A1 (83.61) – A2 (89.67) – A3 (88.45) ***p*** = .042
B1 (86.81) – B2 (91.83) – B3 (88.57) n.s.

(f)

Factor: ***exposure to English***
D.V.: /æ/ percent correct identifications
A1 (61.76) – A2 (55.55) – A3 (46.70) n.s.
B1 (47.79) – B2 (69.38) – B3 (60.47) n.s.

(g)

Factor: ***exposure to English***
D.V.: /ɒ/ percent correct identifications
A1 (50.42) – A2 (57.14) – A3 (61.53) n.s .
B1 (57.14) – B2 (54.08) – B3 (57.14) n.s

(h)

Factor: ***exposure to English***
D.V.: /u/ percent correct identifications
A1 (54.62) – A2 (58.73) – A3 (67.03) n.s.
B1 (54.94) – B2 (74.48) – B3 (61.19) n.s.

(i)

Factor: ***exposure to English***
D.V.: /ʌ/ percent correct identifications
A1 (66.38) – A2 (67.19) – A3 (53.84) n.s.
B1 (65.93) – B2 (61.22) – B3 (65.71) n.s.

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .007. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .007 as the resulting adjusted $p < .05$.

5.3.2.3.3. Effect of dominant L1(s)

Percent correct identification scores obtained by NSs of English and FL learners of English (Catalan dominant speakers, Spanish dominant speakers, and Catalan/Spanish balanced bilinguals) were submitted to Kruskal-Wallis analyses with dominant L1(s) as a factor, yielding significant differences among language groups in the percent scores for /i/ in both *speak* and *tea* (χ^2 15.025, *df* 3, adjusted $p < .05$; χ^2 19.292, *df* 3, adjusted $p < .05$), /ɪ/ in *it* and *this* (χ^2 27.651, *df* 3, adjusted $p < .05$; χ^2 18.812, *df* 3, adjusted $p < .05$), /ɛ/ (χ^2 17.836, *df* 3, adjusted $p < .05$), /æ/ (χ^2 18.348, *df* 3, adjusted $p < .05$), and /u/ (χ^2 23.508, *df* 3, adjusted $p < .05$). Differences in scores for /ɒ/ only approached significance (χ^2 6.629, *df* 3, $p = .085$, adjusted $p > .05$), while for /ʌ/ they were nonsignificant (χ^2 3.664, *df* 3, $p = .300$, adjusted $p > .05$). All the differences took place between the scores obtained by the control group and those of Catalan dominant speakers, Spanish dominant speakers, and Catalan/Spanish balanced bilinguals, according to Mann-Whitney *U* tests (adjusted $p < .05$).

Among the three language subgroups of learners (with no further distinction of age of onset of FL learning and exposure) there were no significant differences. And only the differences in scores for /i/ (*speak*) (*U* 1056, *Z* –1.885, $p = .051$) and /u/ (*U* 1044, *Z* –1.962, $p = .050$, adjusted $p > .05$) between Spanish dominant speakers and Catalan/Spanish balanced bilinguals approached significance. But no specific language group obtained consistently higher correct identification scores than the remaining language groups. All these results are displayed in Figure 5.89 and summarised in Table 5.56.

When the file was split by subject group, no significant differences were found between the three language subgroups. Furthermore, the correct identification scores obtained for the language subgroups within each learner group corroborated the finding of no consistent “advantage” of one language group over the other in the production of English vowel segments (thus, no further figures with those results are presented).

Figure 5.89. Correct identification scores for /i, ɪ, ɛ, æ, ɒ, u, ʌ/. Factor: dominant L1(s).

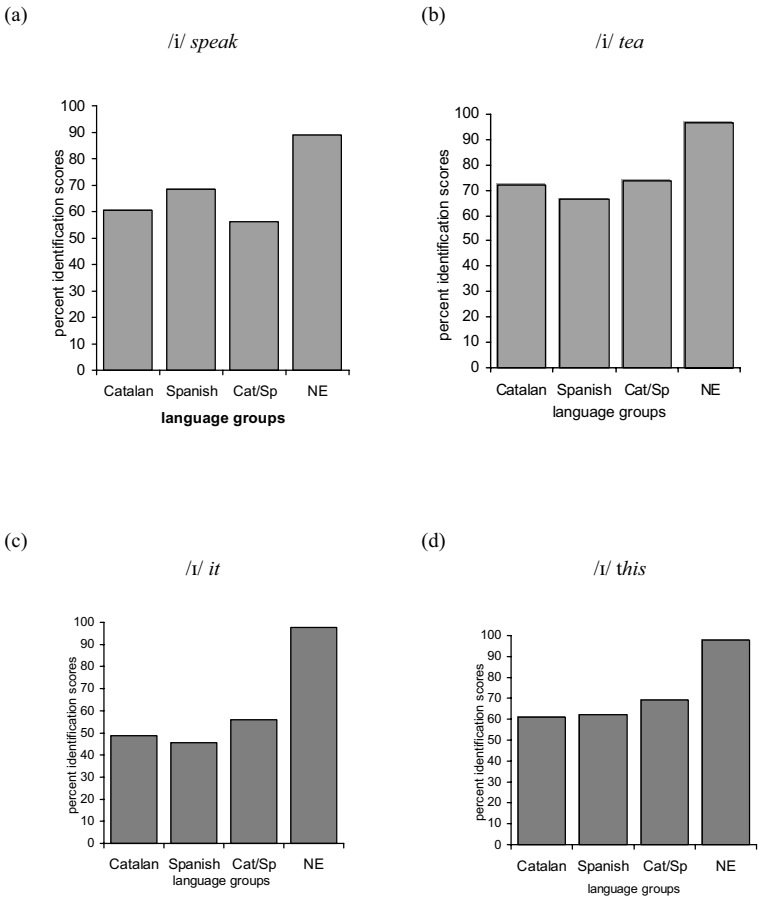
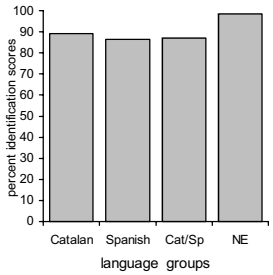


Figure 5.89 (continued)

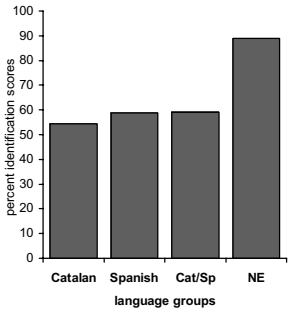
(e)

/ɛ/ *red*, *tests*



(f)

/æ/ *back*, *pad*



(g)

/ɒ/ *box*

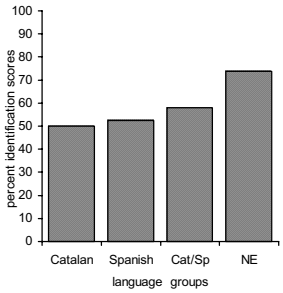
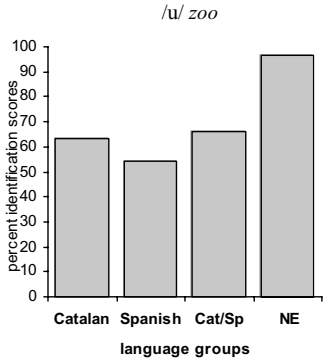


Figure 5.89 (continued)

(h)



(i)

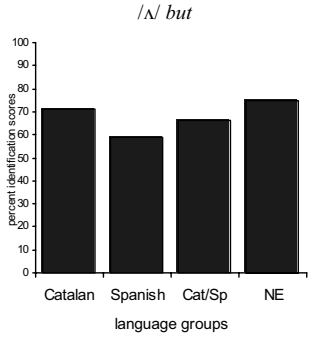


Table 5.56. Summary of comparisons carried out on the vowel percent correct identifications with dominant L1(s) as a factor. Significant comparisons are marked with * ($p < .05$), while nonsignificant group comparisons are displayed as n.s. ($p > .05$). *P* values of marginally significant results ($.05 < p < .10$) are stated as p = .exact significance value. Each group's mean percent correct identification scores for each vowel appear in parentheses.

(a)

Factor: ***dominant L1(s)***
D.V.: /i/ (*speak*) percent correct identifications

Cat (60.50) – Sp (68.33) – C/S (56.16) – NE (89.01)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat – NE*
Sp > C/S ***p*** = .059
Sp < NE ***p*** = .012
C/S < NE*

(b)

Factor: ***dominant L1(s)***
D.V.: /i/ (*tea*) percent correct identifications

Cat (71.84) – Sp (66.40) – C/S (73.00) – NE (96.70)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat – NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(c)

Factor: ***dominant L1(s)***
D.V.: /ɪ/ (*it*) percent correct identifications

Cat (48.73) – Sp (45.55) – C/S (55.96) – NE (97.80)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat – NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(d)

Factor: ***dominant L1(s)***
D.V.: /ɪ/ (*this*) percent correct identifications

Cat (60.92) – Sp (62.16) – C/S (69.27) – NE (97.80)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat – NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(e)

Factor: ***dominant L1(s)***
D.V.: /ɛ/ percent correct identifications

Cat (89.07) – Sp (86.48) – C/S (87.08) – NE (98.35)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat – NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(f)

Factor: ***dominant L1(s)***
D.V.: /æ/ percent correct identifications

Cat (54.41) – Sp (58.68) – C/S (59.88) – NE (89.01)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat – NE*
Sp – C/S n.s.
Sp < NE*
C/S < NE*

(g)

Factor: ***dominant L1(s)***
D.V.: /ɒ/ percent correct identifications

Cat (50.00) – Sp (52.50) – C/S (57.92) – NE (73.62) ***p*** = .085

(h)

Factor: ***dominant L1(s)***
D.V.: /u/ percent correct identifications

Cat (63.44) – Sp (54.44) – C/S (66.34) – NE (96.70)*
Cat – Sp n.s.
Cat – C/S n.s.
Cat – NE*
Sp < C/S ***p*** = .050
Sp < NE*
C/S < NE*

(i)

Factor: ***dominant L1(s)***
D.V.: /ʌ/ percent correct identifications

Cat (71.00) – Sp (59.07) – C/S (66.53) – NE (74.72) n.s.

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .007. Thus, * marks a significant result with adjusted $p < .05$. And marginally significant results should be compared to the alpha level of .007 as the resulting adjusted $p < .05$.

5.3.2.2.4. Effect of gender

As seen in Figure 5.90 below, female subjects' production of English vowels was mostly identified as the intended vowel segments at higher frequency rates than male subjects' vowel productions[103]. Mann-Whitney *U* tests revealed that the differences in identification scores in /ɛ/ and /æ/ between male and female speakers were large enough to be significant (/ɛ/ mean: 90.97 for female Ss vs. 84.06 for male Ss; and /æ/ mean: 67.36 for female Ss vs. 51.75 for male Ss) (*U* 2095.5, *Z* –3.547, adjusted $p < .05$ for /ɛ/; and *U* 1970, *Z* –3.898, adjusted $p < .05$ for /æ/). And differences in scores for /u/ between male and female subjects approached significance (*U* 2155, *Z* –1.815, $p = .070$, adjusted $p > .05$) (Table 5.57).

When the file was split by subject group, none of the gender comparisons reached significance – at the most they approached significance. In spite of this, female subjects still obtained higher correct identification scores than male subjects within each learner group. Like the variable of dominant L1(s), figures and summary tables for each learner group as a function of gender are not presented.

[103] English native speakers were not included in this analysis for the reasons mentioned above (e.g. Section 5.2.2.1.4, footnote 83).

Figure 5.90. Correct identification scores for /i, ɪ, ɛ, æ, ɒ, u , ʌ/. Factor: gender.

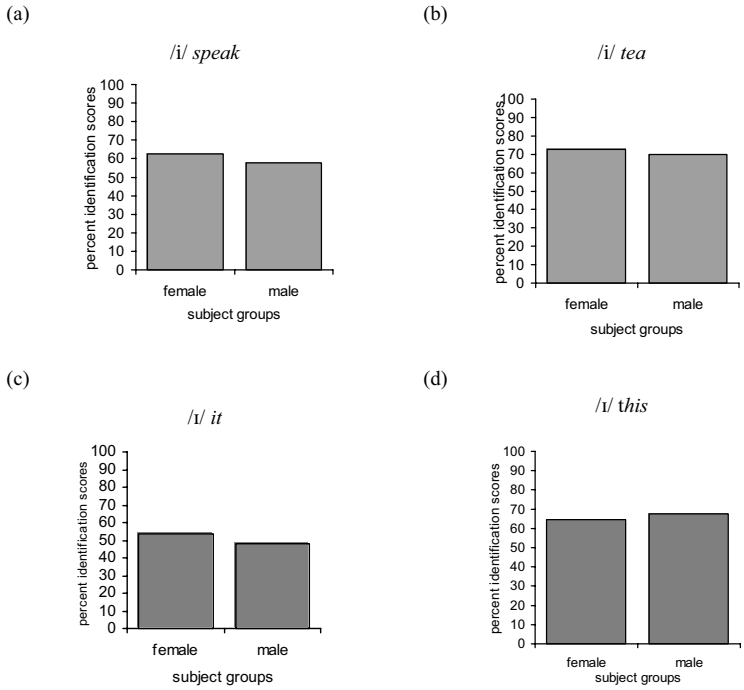
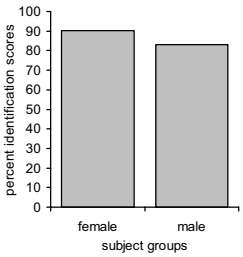


Figure 5.90 (continued)

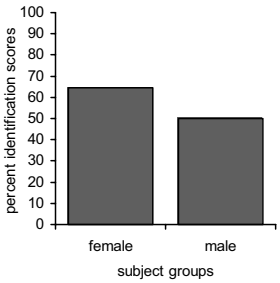
(e)

/ɛ/ *red*, *tests*



(f)

/æ/ *back*, *pad*



(g)

/ɒ/ *box*

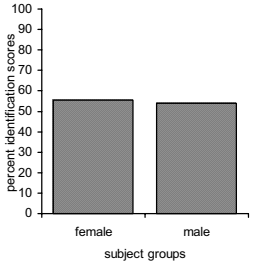
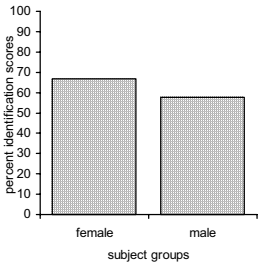


Figure 5.90 (continued)

(h)

/u/ *zoo*



(i)

/ʌ/ *but*

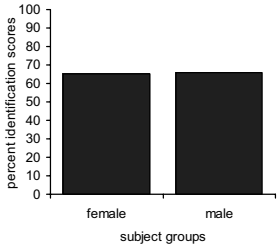


Table 5.57. Summary of comparisons carried out on the vowel percent correct identifications with gender as a factor. Significant comparisons are marked with * ($p <$ .05), while nonsignificant group comparisons are displayed as n.s. ($p >$ .05). *P* values of marginally significant results (.05 < $p <$ .10) are stated as $p =$ .exact significance value. Each group's mean percent correct identification scores for each vowel appear in parentheses.

(a)

Factor: ***gender***
D.V.: /i/ (*speak*) percent correct identifications
Female (62.56) – Male (57.61) n.s.

(b)

Factor: ***gender***
D.V.: /i/ (*tea*) percent correct identifications
Female (72.74) – Male (69.76) n.s.

(c)

Factor: ***gender***
D.V.: /ɪ/ (*it*) percent correct identifications
Female (53.36) – Male (48.33) n.s.

(d)

Factor: ***gender***
D.V.: /ɪ/ (*this*) percent correct identifications
Female (64.53) – Male (67.61) n.s.

(e)

Factor: ***gender***
D.V.: /ɛ/ percent correct identifications
Female (90.22) – Male (82.97)*

(f)

Factor: ***gender***
D.V.: /æ/ percent correct identifications
Female (64.44) – Male (50.00)*

(g)

Factor: ***gender***
D.V.: /ɒ/ percent correct identifications
Female (55.33) – Male (54.04) n.s.

(h)

Factor: ***gender***
D.V.: /u/ percent correct identifications
Female (66.83) – Male (57.61) ***p*** = .070

(i)

Factor: ***gender***
D.V.: /ʌ/ percent correct identifications
Female (65.02) – Male (65.71)*

Note: for a result to be significant at the .05 level, the significance level obtained should be <= .007. Thus, * marks a significant result with adjusted $p <$ .05. And marginally significant results should be compared to the alpha level of .007 as the resulting adjusted $p <$ .05.

5.3.2.3.5. Misidentification patterns

As was the case of the percent correct identification scores, the misidentification scores obtained for each of the target vowel sounds were submitted to Kruskal-Wallis analyses (or Mann-Whitney *U* tests) with onset age of FL learning, exposure to the FL, dominant L1(s), and gender as factors (one factor at a time)[104]. For the most part, differences in misidentifications scores for the seven English vowel segments under study as a function of the four research variables were nonsignificant ($p > .05$). That is, when subjects mispronounced a vowel sound, the substitutions reported were fairly similar regardless of their age of first exposure to English, the amount of exposure to the TL, their dominant L1(s) and gender.

Based on the lack of significant differences in the misidentification scores obtained for the various learner groups, each of the misidentification patterns was pooled into a single score averaged across the 148 learners. This reduction of data, in turn, helped to identify the most frequent substitutions that subjects made for the target sounds.

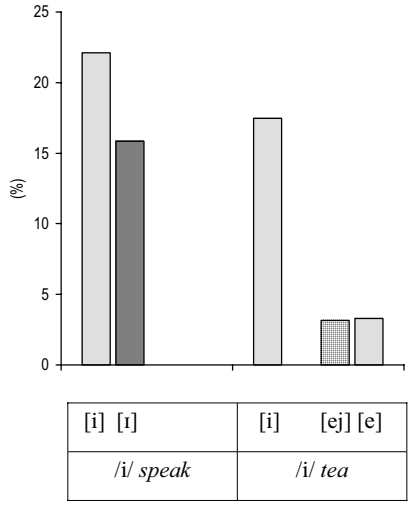
Figures 5.91 – 5.97 below show the percent misidentification scores for each of the seven English vowel segments[105]. These figures further illustrate the extent to which subjects made use of a specific sound substitution for a given target segment. In addition, some variability in percent misidentification scores can be observed. For instance, subjects' productions of /ɪ/ (*it*) were heard as [i] (non-diphthongised) with a frequency rate almost twice as high as that of [ij] (diphthongised) (22% vs. 11.1% for [i] and [ij], respectively). Friedman tests showed that the differences in percent misidentification patterns obtained for each of the vowel segments[106] were significant ($p < .05$). Then, Wilcoxon Signed Ranks Tests showed where the differences occurred.

Therefore, in the case of /i/ (Figure 5.91), [i] (in both words) obtained significantly higher percent misidentification scores than [ɪ] (for *speak*) and [ej] and [e] (for *tea*) (Z –2.666, $p < .05$ for *speak*; Z –7.383, $p < .05$ and Z –6.820, $p < .05$ for *tea*). Differences in scores for [ej] and [e] were nonsignificant (Z –.426, $p > .05$).

---

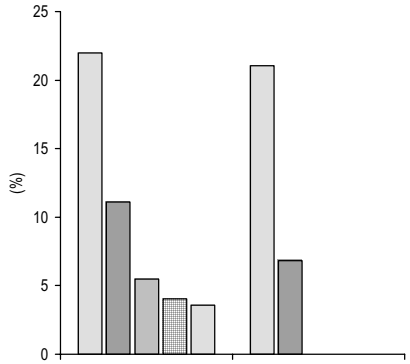
[104] Native English-speaking subjects were not included in these statistical analyses.
[105] As stated above, only substitutions with a frequency of appearance higher than 2% were examined.
[106] Misidentification scores for /ɛ/ were not included in the analysis, as there was only one reported substitute for /ɛ/, namely [e].

Figure 5.91. Misidentification patterns for /i/ (*speak* and *tea*) averaged over all learner subjects.



As for /ɪ/, [i] was the most frequent substitute for both words *it* and *this* (Figure 5.92). [i] also differed significantly from the other reported substitutes. Thus, [i] obtained significantly higher misidentification rates than [ij], [ɛ], [ej], and [e] in *it* (Z –4.564, $p <$ .05; Z –5.469, $p <$ .05; Z –6.454, $p <$ .05; Z –6.562, $p <$ .05); and than [ij] in *this* (Z –6.382, $p <$ .05). In *it*, [ij] was heard at higher frequency rates than [ɛ], [ej], and [e] (Z –3.075, $p <$ .05; Z –3.771, $p <$ .05; Z –4.227, $p <$ .05). The latter three substitutes did not differ significantly in the frequency rates that they were heard.

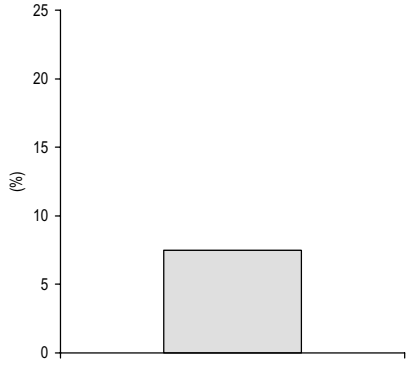
Figure 5.92. Misidentification patterns for /ɪ/ (*it* and *this*) averaged over all learner subjects.



[i] [ij] [ɛ] [ej] [e]	[i] [ij]
/ɪ/ *it*	/ɪ/ *this*

Only one sound substituted for /ɛ/ at a frequency rate higher than 2%: [e] (Figure 5.93). Besides, the target sound /ɛ/ was often identified as such; hence, leaving little room for substitutions.

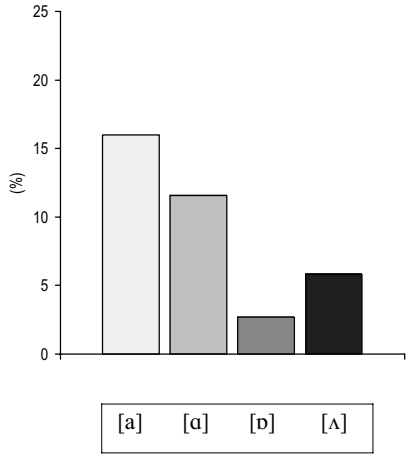
Figure 5.93. Misidentification patterns for /ɛ/ (*red*, *tests*) averaged over all learner subjects.



[e]

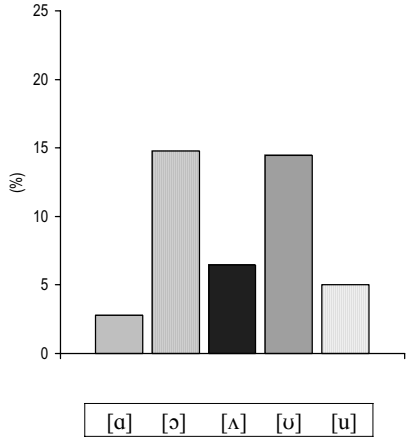
Concerning /æ/ (Figure 5.94), [a] was the most frequent substitute, whose frequency rate significantly differed from the other reported substitutes [ɑ], [ɒ], and [ʌ] (Z –3.357, $p < .05$; Z –8.923, $p < .05$; Z –6.976, $p < .05$). In fact, in all two-sound comparisons, significant differences were found. Therefore, [ɑ] obtained higher misidentification scores than [ɒ] and [ʌ] (Z –7.982, $p < .05$; Z –4.537, $p < .05$). And so did [ɒ] in relation to [ʌ] (Z –3.167, $p < .05$).

Figure 5.94. Misidentification patterns for /æ/ (*back*, *pad*) averaged over all learner subjects.



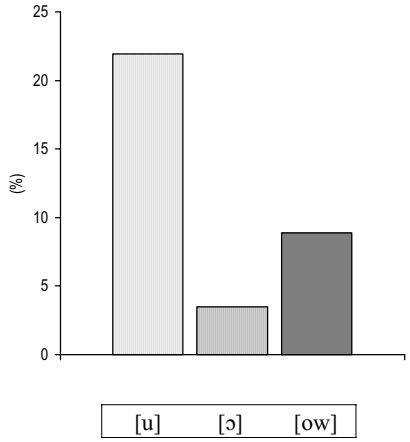
/ɒ/ presented a variety of substitutions (Figure 5.95). At one end the most frequent substitutions were [ɔ] and [ʊ] – which did not differ significantly from each other (Z –.613, $p > .05$). And in the middle point, there were [ʌ] and [u] that did not differ significantly from each other, either (Z –.674, $p > .05$). Thus, [ɑ] was the least frequent substitution, significantly differing from the other four substitutions (Z –7.147, $p < .05$ for [ɑ]-[ɔ]; Z –2.634, $p < .05$ for [ɑ]-[ʌ]; Z –5.477, $p < .05$ for [ɑ]-[ʊ], Z –2.383, $p < .05$ for [ɑ]-[u]. Moreover, [ɔ] obtained significantly higher frequency rates than [ɑ], [ʌ], and [u] (Z –7.147, $p < .05$; Z –4.460, $p < .05$; Z –5.870, $p < .05$). And so did [ʌ] in comparison to [ɑ] and [u] (Z –2.634, $p < .05$; Z –3.506, $p < .05$).

Figure 5.95. Misidentification patterns for /ɒ/ (*box*) averaged over all learner subjects.



Three misidentification patterns were found for /u/, i.e. [ʉ], [ɔ], and [ow], being [ʉ] the most frequent substitute (Figure 5.96). In all cases, the comparisons carried out on the misidentification scores resulted in significant differences. Thus, [ʉ] obtained significantly higher misidentification rates than [ɔ] and [ow] (Z –7.589, $p < .05$; Z –5.223, $p < .05$), and [ow], in turn, obtained higher misidentification rates than [ow] (Z –3.840, $p < .05$).

Figure 5.96. Misidentification patterns for /u/ (*zoo*) averaged over all learner subjects.



Finally, [æ] was reported to be the most frequent substitute for /ʌ/ (Figure 5.97), obtaining significantly higher rates than the other misidentification patterns: [a], [ɑ], [ɒ], and [ɔ] (Z –2.754, $p < .05$; Z –4.348, $p < .05$; Z –2.561, $p < .05$; Z –2.832, $p < .05$). In addition, [a] and [ɒ] also obtained significantly higher scores than [ɑ] (Z –3.379, $p < .05$; Z –2.329, $p < .05$).

Figure 5.97. Misidentification patterns for /ʌ/ (*but*) averaged over all learner subjects.