# Assessing the Effect of Perceptual Training on L2 Vowel Identification, Generalization and Long-term Effects

Angélica Carlet
Universitat Internacional de Catalunya

Juli Cebrian
Universitat Autònoma de Barcelona

**Abstract**
This paper assessed two high variability phonetic training methods aimed at improving the perception and production of English vowels by Spanish/Catalan speakers. Fifty-four L2 learners of English were assigned to one of three groups: forced-choice identification (ID) training, AX categorical discrimination (DIS) training, and control group (CG). Participants' identification and production of English vowels was assessed before training, after training and two months later. Both trained groups outperformed the CG at posttest and showed evidence of generalization and retention of learning. However, the ID trainees showed greater improvement in perception and significant gain in production, pointing to a potential superiority of this method for vowel learning. These results have implications for future research on phonetic training and practical applications for the teaching of pronunciation.

## 1. Introduction

The acquisition of target second language (L2) sounds can be challenging for the L2 learner due to the interplay of many factors including onset age of learning, length of residence in the target-language country, amount of L2 exposure, amount of L1 and L2 use, learner motivation and aptitude, and linguistic factors like typological relatedness or the role of orthography

Anne Mette Nyvad, Michaela Hejná, Anders Højen, Anna Bothe Jespersen & Mette Hjortshøj Sørensen (Eds.), ***A Sound Approach to Language Matters – In Honor of Ocke-Schwen Bohn*** (pp. 91-119). Dept. of English, School of Communication & Culture, Aarhus University.

(Piske, MacKay & Flege, 2001; Bohn & Munro, 2007). This difficulty is clearly related to the effect of existing L1 phonetic categories[1] and the L2 learners' failure to perceive target L2 sounds accurately, as proposed by L2 speech models such as Flege's (1995a, 2003) Speech Learning Model (SLM), Kuhl and Iverson's (1995) Native Language Model (NLM), and Best and Tyler's (2007) Perceptual Assimilation Model (PAM-L2), among others. According to these models, given enough input and experience, learners may succeed in establishing long-term memory representations for target L2 sounds, separate from pre-existing L1 categories.

The present study is set in an instructional context, that is, learning English as a foreign language in the learners' home country. This setting is characterized by limited exposure to the target language outside the classroom (Muñoz, 2008; Saito, 2015). This scenario may be problematic for accurate second language learning, since extensive exposure to the target language is crucial to develop the ability to distinguish native from non-native sounds (Flege, 1991; Ingram & Park, 1997), a pre-requisite for accurate L2 category formation (e.g., Flege, Bohn & Jang, 1997; Flege, 1995a). Against this background, a possible source of specialized target language input can be found in phonetic training, which aims at directing L2 learners' attention to challenging features or contrasts present in the target language by means of specialized perceptual or pronunciation tasks that generally include corrective feedback (Cebrian & Carlet, 2014). There is evidence that a short training regime may have the same outcome as a prolonged period of instruction, and that training is effective even for learners at different levels of proficiency. Pereira (2014) reported that a group of Chilean learners of English who completed a six-week perceptual training regime were able to improve their perception of English vowels to a similar extent as another group of Chilean learners who had undergone three years of formal instruction. Iverson, Pinet and Evans (2012) explored whether training was equally effective for different settings and levels of proficiency. Beginner and intermediate French learners underwent a vowel identification training regime and were tested on the identification, discrimination and production of 14 English vowels and diphthongs. Both groups showed a slight effect of training on discrimination ability, as well as significantly improved their identification and production as a result of training.

[1] Phonetic categories are defined as "the distribution of acoustic tokens which together are perceived as mapping to a phoneme in the listener's inventory" (Earle & Myers, 2014, p. 1192).

There is thus evidence from a considerable amount of studies that phonetic training can be beneficial for different L1-L2 language combinations and different target structures, particularly L2 consonants and vowels (Cebrian & Carlet, 2014; Iverson & Evans, 2007, 2009; Lacabex, García-Lecumberri & Cooke, 2008; Lengeris, 2008; Nishi & Kewley-Port, 2007; Nobre-Oliveira, 2007; Rato, 2014; Thomson, 2012). A number of laboratory training studies have adopted successfully what is known as a high variability phonetic training approach (HPVT), which incorporates multiple stimuli involving a variety of speakers, tokens, phonetic contexts, etc., in an attempt to replicate the variability that characterizes L2 input in a natural environment (Logan, Lively & Pisoni, 1991; Lively, Pisoni & Logan, 1993; see section 1.1). It has been argued that training is truly effective if its effect goes beyond improvement on the trained structures from pretest to posttest, that is, if improvement generalizes to untrained stimuli such as new voices, new items or new modalities (Logan & Pruitt, 1995; Flege, 1995b; Bradlow, 2008). In addition, the efficacy of phonetic training is demonstrated further when the observed improvement is still found some time after training has ended, that is, if learning is retained beyond the training period. According to Logan and Pruitt (1995), generalization and retention provides evidence that robust learning has occurred. This study examines the effect of high variability perceptual training on L2 vowel perception and production and compares the effectiveness of two types of perceptual tasks, identification and discrimination, on the ability to identify and produce L2 sounds. In addition, the study also compares the two perceptual methods on the extent to which the potential improvement generalizes to untrained structures, and is retained after a two-month interval.

### 1.1. Perceptual training studies on vowels

The learnability of vowels through laboratory training has been investigated extensively in the last few decades (Aliaga-García & Mora, 2009; Cebrian & Carlet, 2014; Iverson & Evans, 2007, 2009; Lacabex et al., 2008; Lambacher, Martens, Kakehi, Marasinghe & Molholt, 2005; Lengeris, 2008; Nishi & Kewley-Port, 2007; Nobre-Oliveira, 2007; Rato, 2014; Rato & Rauber, 2015; Thomson, 2012; Wang & Munro, 2004; among others). For instance, in an HVPT study, 26 Mandarin Chinese speakers were trained to improve the perception of 10 English vowels produced in a post labial stop context (Thomson, 2012). After eight short identification training sessions, learners' ability to identify the English vowels improved

significantly and also generalized to a velar stop context. Moreover, the improvement obtained after training was retained one month after training completion. In fact, several studies have also reported successful retention of learning after periods of time ranging from one to twelve months (Rato, 2014; Wang & Munro, 2004; Nishi & Kewley-Port, 2007), which confirms the robustness of the training procedure and the relevance of phonetic training as an L2 learning tool (Logan & Pruitt, 1995).

Aliaga-García and Mora (2009) investigated the effect of HVPT in a study involving Spanish/Catalan learners of English and found a positive effect of HVPT on the identification and, to a lesser extent, production of English initial stops. Training also improved vowel perception; however, no positive effect of training on vowel production was observed. In a later study, Aliaga-García, Mora and Cerviño-Povedano (2011) found that improvement in L2 vowel perception varied as a function of phonological short-term memory capacity. Further, in a short-term perceptual training study involving Spanish/Catalan speakers, Cebrian and Carlet (2014) assessed the effect of a three-week HVPT regime (four 45-minute sessions) consisting of a variety of perceptual tasks on the perception of two vowel pairs (/**i:**/-/ɪ/ and /æ/-/ʌ/, as well as two consonant contrasts) by advanced learners of English. They found a positive effect of training for a subset of the target vowels, namely /**i:**/ and /ʌ/, and partial generalization effects. Finally, Rato (2014) and Rato & Rauber (2015) reported both generalization and retention of learning after a training regime that combined identification and discrimination tasks. These studies, however, combined different perceptual training tasks in the same training regime, so it is not possible to evaluate what the relative contribution of the different tasks may have been. The present study tries to contrast and evaluate the effectiveness of each type of task.

### 1.2 Perceptual training tasks

Perceptual training studies often make use of discrimination or identification tasks. Even though early findings with stop consonants revealed the efficacy of discrimination (DIS) tasks in modifying learners' categorical perception of these sounds (Pisoni, Aslin, Perey & Hennessy, 1982; McClaskey, Pisoni & Carrell, 1983), the efficacy of identification (ID) training has been said to be superior to discrimination training as an L2 training tool (Jamieson & Morosan, 1986; Logan & Pruitt, 1995, among others). Strange and Dittmann (1984) found that Japanese learners of English improved their identification and discrimination of English /r/-

/l/ after undergoing auditory discrimination training involving synthetic stimuli. However, this improvement did not generalize to novel and natural stimuli. By contrast, identification tasks have been found to promote generalization of learning (Jamieson & Morosan, 1986; Logan et al., 1991). It is possible that DIS tasks promote within-category sensitivity and tap into lower levels of phonological encoding that are not greatly affected by language experience, while ID tasks may enhance between-category sensitivity and involve higher levels of phonological encoding more relevant for L2 categorization (Jamieson & Morosan, 1986; Logan & Pruitt, 1995; Iverson et al., 2012). Still it has been proposed that both ID and categorical DIS may affect similar levels of processing (Flege, 2003; Højen & Flege, 2006) and hence equally promote categorization of L2 sounds (Polka, 1992).

Few prior studies have compared the efficacy of ID and categorical DIS tasks incorporating highly variable stimuli in the same study (Flege, 1995b; Wayland & Li, 2008; Nozawa, 2015, Shinohara & Iverson, 2018). Flege (1995b) assessed the efficacy of both types of task (two-alternative forced-choice identification task and categorical same/different discrimination task) in a single HVPT study aimed at training Mandarin learners of English to identify final /d/ and /t/. Identification scores after seven training sessions showed that the two trained groups outperformed the controls at post-test and showed generalization of knowledge and long term effects. These findings pointed to the efficacy and robustness of both training methods and challenged the general claim that ID training is superior to discrimination training. Wayland and Li (2008) trained Chinese and English listeners to discriminate Thai tone contrasts by means of ID and DIS tasks. The findings revealed that both ID and DIS training procedures were similarly effective in enhancing listeners' discrimination of Thai tone contrasts and that the Chinese group outperformed the English group. Thus, the authors concluded that both methods were equally effective in improving tone perception and that the prior experience with a tone language explained the Chinese participants' advantage.

On the other hand, Nozawa (2015) compared the effect of ID and categorial ABX DIS training on Japanese learners' identification of English coda nasals and vowels in a small scale study involving two training sessions. While Nozawa found that both methods promoted significant gains regarding the final nasals, the ID method was found to be superior to ABX DIS for training vowels. A recent study compared the efficacy of the DIS and ID tasks further by evaluating their effect on the perception

and production of the /r/-/l/ contrast by Japanese adult learners of English (Shinohara & Iverson, 2018). L2 learners were assessed on identification, auditory discrimination, category discrimination, and /r-l/ production at three times (pretest/midtest/post-test). Experimental groups were trained with both tasks in a different order. Their results after a 10 session regime showed that both training methods improved Japanese speakers' perception and production of English /r-l/ to a similar extent. In summary, more recent studies comparing ID and categorical DIS tasks have provided comparable results for both methods, particularly for training consonants. To our knowledge, only the study by Nozawa (2015) investigated vowels, showing a greater effect of ID in this case. This study explores the effects of these two methods further by contrasting their effect on L2 vowel perception and production. The main questions the present study aims to address are:

- Which type of training (ID or DIS) is more efficient in promoting improvement on the perception of L2 vowel sounds by Spanish/Catalan bilinguals?
- Which type of training (ID or DIS) is more efficient in promoting improvement on the production of L2 vowel sounds by Spanish/Catalan bilinguals?
- Which type of training (ID or DIS) is more efficient in promoting generalization and long-term effects?

Assuming that both training methods (ID and categorical DIS) tap into similar levels of processing (Flege, 2003; Højen & Flege, 2006) and also promote L2 categorization (Polka, 1992), it is hypothesized that both methods will be equally effective in improving learners' perception after training as well as promoting generalization of learning and retention effects, in accordance with Flege (1995b). Moreover, perceptual training with no focus on production may lead to production gains, even if to a lesser extent than the perceptual gains (Rato & Rauber, 2015; Rochet, 1995; Bradlow, 2008; Hardison, 2004; Iverson & Evans, 2009; Thomson, 2012; Pereira, 2014).

## 2. Methods

### 2.1. Participants

Fifty-four learners of English as an L2 took part in a 10-weeek-long regime and were assigned to one of three groups: 1) forced-choice identification

training (ID, *N*=20), b) AX categorical discrimination training (DIS, *N*=18), or c) control group with no perceptual training (CG, *N*=16).[2] The L2 learners were Catalan/Spanish bilinguals, with a mean age of 19.7, and an initial age of EFL learning of 5.75 years. All subjects were second-year undergraduate students in English Studies at the *Universitat Autònoma de Barcelona (UAB)* enrolled in an introductory phonetics course. The learners' level of English ranged from a B2 to a C1 level on the *Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (CEFRL) (Council of Europe, 2001), with limited experience in an English-speaking country (average: two weeks) and no self-reported hearing impairments. Participants received course credit for their participation.

**2.2. Target sounds and stimuli**

The target sounds were the five standard Southern British English (SBE) vowels /iː ɪ æ ʌ ɜː/, which are challenging for native speakers of Spanish/Catalan (Cebrian, 2006; Cebrian, Mora & Aliaga-García, 2011). The stimuli consisted of unmodified CVC nonsense words and real words elicited from ten native speakers of standard Southern British English (SBE) (five females and five males, mean age 27.8, range 23-39). The target vowels were always preceded and followed by obstruent consonants The words were elicited by means of the carrier sentence: *I say "word", I say "word" now, I say "word" again.* In order to ensure the desired pronunciation of the nonsense words, the phrase *It rhymes with "real word",* was added at the beginning (e.g., *It rhymes with give, I say "tiv"* ....). Recordings took place in a soundproof booth at the speech laboratory at University College London, UK, and each word was recorded three times. The recordings were carried out using *Cool edit 2000* software, a *Rode NT-1AX* microphone, *Edirol UA25* audio interface and were digitized at a 44.1 kHz sampling rate and 16 bit quantification.

**2.3. Training stimuli**

Training words consisted of nonsense words, so as to eliminate a potential word familiarity effect, given that the use of real words has been found to affect the accuracy and speed of word processing (Grosjean, 1980). These words were obtained from four of the SBE native speakers (two males and two females) with the objective to provide variability, as is characteristic

[2] Originally there were 20 participants in each group, e.g. at pretest, but a few learners did not complete all the training sessions and were discarded.

of HVPT. There were twelve words per target vowel (/iː ɪ æ ʌ ɜː/), plus six words for two additional vowels (/e/ and /ɑː/). The latter two were included to be contrasted with /ɜː/. Thus there were a total of 288 training stimuli (72 nonsense words x 4 talkers). The same stimuli were used in the identification and discrimination training tasks, as explained below. A list of the perceptual training stimuli can be seen in Appendix 1.

**2.4. Testing stimuli**

Testing stimuli consisted of a subset of the non-words used at the training phase and involved 30 words (i.e. 5 target vowels x 6 words) of CVC nonsense words produced by 2 novel talkers (one male and one female), that is, different from training talkers, resulting in 60 testing stimuli. Since stimuli from these talkers were not used in the training corpus, testing already examined generalization to new talkers. In addition, 7 non-words were included as practice tokens in order to guarantee that the task procedure was understood and eight non-words involving the vowels /e/ and /ɑː/ were included as testing fillers. Additionally, 20 CVC real word stimuli and 20 novel non-word stimuli produced by two familiar talkers (i.e. two of the four training talkers) tested generalization to real words and to novel untrained non-words, respectively (5 vowels x 2 words x 2 talkers).

**2.5. Procedure**

Participants were assessed at three testing times (pre-test, post-test and delayed post-test) by means of the same perception and production tests. The perceptual tests consisted of two 7-alternative forced-choice vowel identification tasks (nonsense and real words) involving stimuli produced by different talkers from those used in the training phase. After training, generalization to new talkers and new words was also assessed by means of the same type of identification tasks. The response alternatives consisted of a phonetic symbol together with two common words representing each sound, specifically: */æ/ ash, mass*; */ʌ/ sun, thus*; */ɪ/ fish, his*; */iː/ cheese, leaf*; */ɜː/ earth, first*; /e/ *less, west*; */ɑː/ arm, palm*. Learners' L2 production was elicited by means of a picture naming task before and after training (pre-test and post-test). Participants were asked to name 27 different pictures and repeat the word twice. The 27 test words included the 10 real words containing the target vowel sounds examined between obstruent consonants.[3] A list of the production words can be seen in Appendix 1.

[3] This study is part of a larger scale study, which investigated the effect of HVPT on both consonants and vowel sounds.

Training for the experimental groups consisted of five 30-minute sessions over a 10 week-period and it was administered using TP software (Rauber, Rato, Kluge & Santos, 2011). An approximate study timeline is shown in Table 1. The DIS group was trained by means of AX discrimination tasks with immediate feedback. Participants responded by clicking on "same" or "different". "Different" trials involved the two high-front vowels (/iː-ɪ/), the two low vowels (/æ-ʌ/) or the central vowel /ɜː/ combined with either /e/ or /ɑː/. Each pair was presented in the two possible orders in the same session (/æ-ʌ/, /ʌ-æ/), and in six different talker combinations over the course of the five sessions. There were 288 trials per training session. The ID group was trained by means of a 7-alternative forced-choice identification task with immediate feedback. The training tasks were specifically designed so as to ensure that both groups were exposed to the exact same set of stimuli through training. Thus, the ID tasks consisted of 576 trials per training session, involving the same stimuli presented in a discrimination session (that is, 288 trials involving a pair of stimuli each). Training for the control group was designed to provide the same amount of L2 instruction as the other groups without specific training. Thus, after the pretest, the controls performed five transcription practice sessions using an online platform, *The web transcription tool* (Cooke, García-Lecumberri, Maidment & Ericsson, 2005). Testing and training took place at the Speech Laboratory at UAB.

| | |
|---|---|
| WEEK 1 | Production pre-test (real words) |
| WEEK 2 | Identification pre-tests – non-words / real words |
| WEEK 3 | Training session 1 (ID / DIS) – non-words |
| WEEK 4 | Training session 2 (ID / DIS) – non-words |
| WEEK 5 | Training session 3 (ID / DIS) – non-words |
| WEEK 6 | Training session 4 (ID / DIS) – non-words |
| WEEK 7 | Training session 5 (ID / DIS) – non-words |
| WEEK 8 | Production post-test + Identification post-test |
| WEEK 9 | Generalization test (new non-words) |
| WEEK 10 (2 months later) | Retention test: Identification tests |

Table 1. Study design and approximate timeline

### 2.6. Analysis

The percent correct identification for each sound by participant and group were calculated for each testing phase (pre-test, post-test, generalization and retention test). The L2 production data was analyzed by means of native English speaker judgments. Four Southern British English speakers were asked first to identify the sound they heard and then to rate it on a 9-point Likert scale, where 1 meant "hard to identify as the selected sound" and 9 "easy to identify as the selected sound".

## 3. Results

### 3.1. L2 vowel perception

Correct identification scores at pre-test and at post-test were calculated for the two groups trained on vowels (ID, DIS) and the control group, and are shown in Table 2 below. Importantly, the groups did not differ statistically at pre-test ($F(2,51=.416, p>.05)$. Therefore, a measure of gain (understood as the difference between posttest and pretest) was calculated (see Figure 1) and will be used for further analyses. Since testing stimuli and training stimuli were from different talkers, the comparison between pretest and posttest scores already examines generalization to new talkers.

| | CONTROL | | DIS | | ID | |
|---|---|---|---|---|---|---|
| Non-words | % | SD | % | SD | % | *SD* |
| PRE | 54.1 | 9.9 | 55.5 | 6.5 | 52.9 | 9.5 |
| POST | 57.8 | 10.2 | 65.3 | 9.7 | 79.1 | 13.3 |

Table 2. Percent correct identification at pretest and posttest per group (non-words).

As shown in Table 2, the three groups had similarly low scores at pretest and performed numerically better at post-test. This is particularly evident in the case of the ID group, whose results rose about 26 percentage points from 52.9% to 79.1% correct identification. Improvement was also observed with the DIS group (9.8 percentage points increase). The numerical improvement obtained by the control group is smaller (3.7 increase) and may reflect the influence of the English phonetics course participants were enrolled in, or simply the result of general exposure to English in this and other courses between the pre-test and the post-test

phases. The gain scores were submitted to a generalized linear mixed-effects model (GLMM), with group (ID, DIS and CG) as the fixed effect and participants as a random effect. The analysis revealed a significant main effect of group $F$ (2,51)=61.288, $p$<.001). The group effect is related to the fact that the control group performed differently from the experimental groups. In fact, sequential Bonferroni pairwise comparisons confirmed that the two experimental groups outperformed the controls on the overall identification of L2 vowels ($p$<.01 for the ID group and $p$<.05 for the DIS group). Moreover, the ID group outperformed the DIS group ($p$<.01).

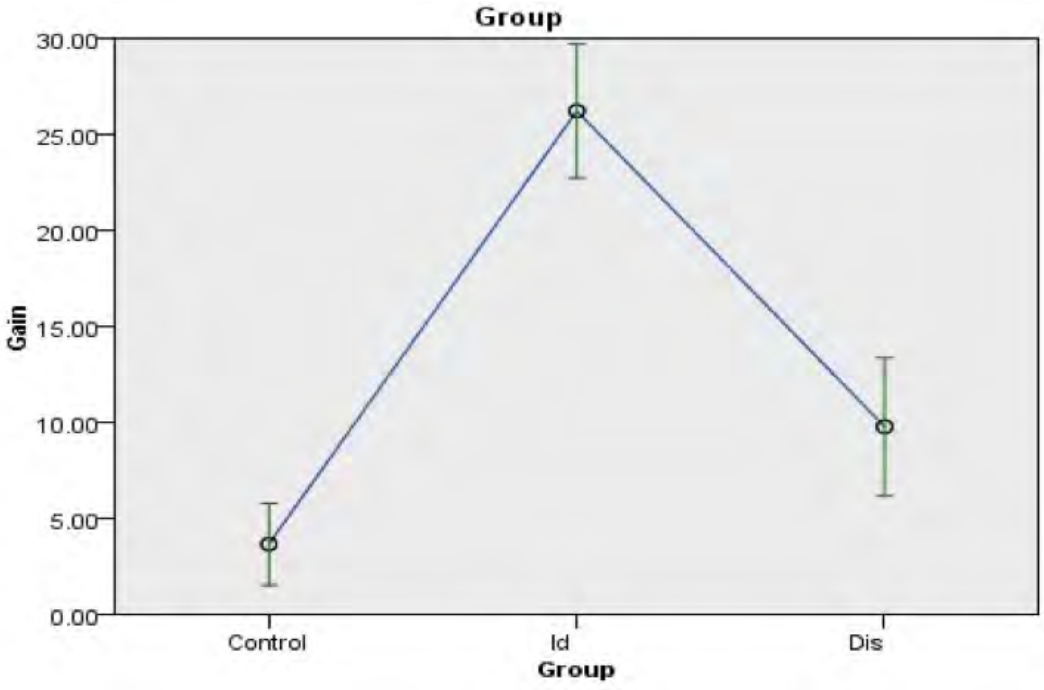


Figure 1. Identification gain (increase in correct identification percentage points) from pre to post-test per group for non-words.

Table 3 shows the mean identification scores at pre and post-test for each individual vowel for each group. It is interesting to note that some vowels seemed to improve more than others. At pre-test, on the whole all groups had the greatest difficulty identifying /æ/ and /ɜː/, followed by /ʌ/, while /iː/ and /ɪ/ were more accurately identified. The ID group is the group that improved the most, and also the one that obtained more comparable results across vowels at post-test (76-83%, compared to 51-81% for DIS). Misidentification errors generally involved /iː/ - /ɪ/ and /æ /- /ʌ/ confusions, while /ɜː/ was most often misheard as /ʌ/ or /e/. The improvement

seen with the control group, mostly for the sounds /ɪ/ and /ʌ/, may be the result of the formal phonetics instruction and the consequent phonological awareness about English sounds. Generally, the ID trainees obtained numerically higher gain scores than the DIS trainees, who in turn also seemed to outperform the CG, in line with the global results across vowels previously described.

| | CONTROL | | DIS | | ID | |
|---|---|---|---|---|---|---|
| SOUND | PRE | POST | PRE | POST | PRE | POST |
| /æ/ | 39.8 | 40.8 | 42.8 | 50.6 | 31.0 | 77.7 |
| | (17.6) | (19.9) | (18.7) | (25.0) | (21.4) | (20.1) |
| /ʌ/ | 55.2 | 62.0 | 53.4 | 66.2 | 53.5 | 75.6 |
| | (22.7) | (21.0) | (12.2) | (18.6) | (20.6) | (18.3) |
| /iː/ | 67.4 | 64.5 | 61.1 | 60.4 | 65.8 | 77.9 |
| | (15.6) | (19.2) | (13.3) | (16.1) | (12.9) | (13.5) |
| /ɪ/ | 69.0 | 80.2 | 72.0 | 81.0 | 75.0 | 82.9 |
| | (15.7) | (14.2) | (15.7) | (14.4) | (14.3) | (13.6) |
| /ɜː/ | 39.1 | 41.4 | 47.9 | 68.1 | 39.0 | 81.5 |
| | (21.0) | (19.6) | (23.7) | (20.8) | (17.8) | (17.0) |

Table 3. Percent correct identification at pretest and posttest for each individual vowel per group (non-words; standard deviations are given in parentheses).

### 3.2 L2 vowel production

Production was assessed at pre-test and at post-test and was analysed by means of native speaker judgments, following Munro (2008), among others, who advocate for the use of listeners' ratings as the most appropriate method of assessment of L2 speech: "From the standpoint of communication, there is no useful way to assess accentedness […] except through listener responses of some sort" (p. 200). In the present study, twelve native English speakers performed a series of rating tasks that included a subset of all the stimuli so that each stimulus was evaluated by four different native English listeners. Seven identification tests with category goodness ratings were created. The rating scale ranged from 1 (difficult to recognize as the selected sound) to 9 (easy to identify as the selected sound). A reliability analysis using an intra-class correlation coefficient (ICC) with a level of "absolute agreement" was conducted on the rating scores. The results revealed a robust inter-rater agreement in all

cases, as Cronbach's alpha values ranged from α = .741 to α = .905. Thus, the median rating score for each participant and group at pre-test and post-test was calculated (see Table 4). The production gain scores were obtained by subtracting the pre-test scores from the post-test scores (Figure 2) and were further submitted to statistical analysis.

| | CONTROL | | DIS | | ID | |
|---|---|---|---|---|---|---|
| L2 Production | Median | *SD* | Median | *SD* | Median | *SD* |
| PRE | 4.5 | 1.4 | 4.7 | 0.8 | 4.6 | 1.5 |
| POST | 4.2 | 0.9 | 5.1 | 1.1 | 5.2 | 1.1 |

Table 4. Median rating for vowel production at pre-test and post-test per group.

As shown in Table 4, the ratings obtained by the control group showed no improvement from pre-test to post-test. The training groups, on the other hand, were given higher ratings after training. More specifically, the DIS group's median scores improved by 0.4 and the ID improved by 0.6. The gain scores for L2 vowel production were submitted to a GLMM, with group (ID, DIS, CG) as fixed effect and participants as random effect.

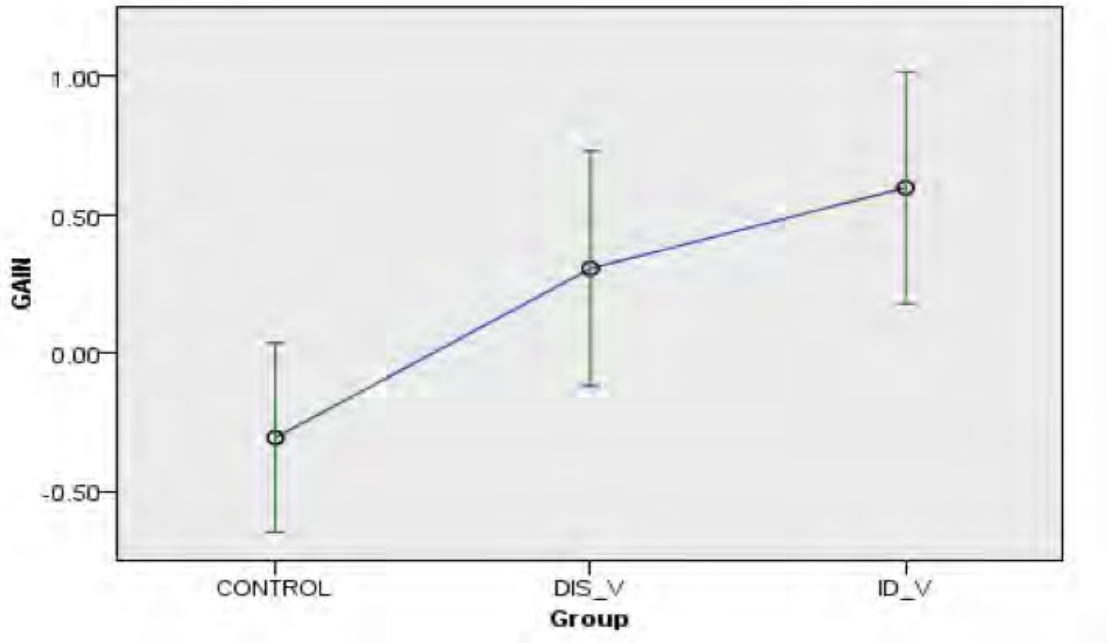


Figure 2. Production improvement from pre to post-test per group (difference between the ratings obtained at posttest and at pretest).

The results yielded a significant main effect of group ($F(2, 50)=6.13, p<.01$), and pairwise comparisons with a sequential Bonferroni correction revealed

that only the ID group significantly outperformed the controls ($p<.01$). The DIS group was marginally significantly better than CG ($p=.057$). Moreover, the two experimental groups didn't differ in performance ($p>.05$), showing a tendency towards a better performance at posttest for the DIS group too. These results suggest that the perceptual training was not only efficient in improving the learners' perception of vowel sounds, but also appeared to modify learners' production as perceived by native English speakers, particularly in the case of the ID group.

With respect to the results obtained per vowel, as was observed for perception, some vowels seemed to yield better results than others and improved to different degrees (see Table 5). No improvement was observed in the production of any of the vowels by the control group. The DIS group improved mostly in the production of vowel /æ/, while the ID group showed some improvement with all the vowels. The vowel that obtained the lowest ratings at the outset was /ʌ/, followed by the vowels /ɪ/ and /æ/. The two highest rated vowels were /iː/ and /ɜː/. These results differ from the perception results mostly regarding two sounds: /ɜː/, which was comparatively less successfully identified than other vowels, and /ɪ/, which was better perceived than other vowels. Both /ʌ/ and /æ/ seemed to pose difficulties to learners both in perception and production and the tense sound /iː/ was the least challenging, particularly in production.

| | CONTROL | | DIS | | ID | |
|---|---|---|---|---|---|---|
| SOUND | PRE | POST | PRE | POST | PRE | POST |
| /æ/ | 4.7 | 4.1 | 4.6 | 5.8 | 4.0 | 4.5 |
| | (2.7) | (1.8) | (1.8) | (2.8) | (2.5) | (2.1) |
| /ʌ/ | 2.0 | 1.8 | 3.3 | 3.3 | 2.3 | 3.4 |
| | (2.2) | (2.8) | (2.6) | (2.2) | (1.7) | (3.2) |
| /iː/ | 5.1 | 4.9 | 5.8 | 6.1 | 6.3 | 6.6 |
| | (1.7) | (3.2) | (2.1) | (2.6) | (1.7) | (3.2) |
| /ɪ/ | 4.8 | 4.4 | 3.9 | 3.9 | 4.7 | 5.1 |
| | (2.5) | (2.9) | (2.5) | (2.9) | (2.7) | (2.2) |
| /ɜː/ | 6.1 | 5.9 | 6.3 | 6.4 | 5.9 | 6.5 |
| | (1.8) | (2.8) | (2.5) | (2.1) | (2.8) | (1.6) |

Table 5. Median ratings obtained for each vowel per group (standard deviations are given in parentheses).

### 3.3 Generalization effects

As previously mentioned, the main results provide evidence of generalization to novel talkers, since testing and training talkers differed. Another type of generalization investigated in this study was generalization to new items (i.e. novel non-words and real words).

#### 3.3.1 Generalization to novel non-words

In order to assess the degree to which the effects of training generalized to novel items (i.e. novel CVC non-words) produced by familiar talkers (talkers heard at the training phase), a further test was administered a week after the post-test took place. The scores of the generalization test are contrasted with both the pre-test scores and post-test scores. Generalization is considered to take place if the generalization results are as high as, or higher than, the post-test scores, and differ from pre-test results. The percentage correct identification at pre-test, post-test and generalization test by the two experimental groups (ID, DIS) and the CG are shown in Table 6. It can be observed that the three groups maintained, or even increased, their vowel identification scores from post-test to generalization test. The fact that the CG group's identification scores in the novel word generalization test were higher than at pre and post-test (68% vs. 54% and 58%, respectively) may be related to the formal instruction received. Alternatively, it is possible that these words or talkers posed fewer problems to the learners. Still, the CG's scores were lower than those obtained by the trainees.

| | CONTROL | | DIS | | ID | |
|---|---|---|---|---|---|---|
| | % | *SD* | % | *SD* | % | *SD* |
| PRE | 54.1 | 9.9 | 55.5 | 6.5 | 52.9 | 9.5 |
| POST | 57.8 | 10.2 | 65.3 | 9.7 | 79.1 | 13.3 |
| GEN WORDS | 68.4 | 12.4 | 75.9 | 8.3 | 80.4 | 9.8 |

Table 6. Percent correct identification at pre-test, post-test and generalization to novel non-words per group.

The results of the GLMM in this case showed a significant effect of group, $F(2,153)=13.977$, $p<.001$, and Bonferroni pairwise comparisons confirmed that both experimental groups outperformed the controls in the perception of target vowels in novel non-words ($p<.001$ for the ID group and $p<.01$ for the DIS group). In order to further explore the results for each experimental

group, GLMM analyses were conducted on the percentage scores obtained by each trained group at the three different tests. Regarding the ID group, the results yielded a significant effect of test ($F(2, 57)=50.42$, $p<.001$), confirming that the ID group performed significantly better after the training than at pre-test. Furthermore, pairwise comparisons with a Bonferroni adjustment confirmed that the ID's pre-test results differed from both the post-test and the generalization test results ($p<.001$). Conversely, post-test and generalization results did not differ significantly. Thus, generalization to novel words was observed for the ID trainees. Regarding the DIS group, the results also revealed a significant test effect ($F(2, 51)=33.693$, $p<.001$) and sequential Bonferroni pairwise comparisons confirmed that the pre-test scores significantly differed from both the post-test and the generalization scores ($p<.01$). Interestingly, the generalization scores were significantly higher than the post-test results ($p<.01$) for the DIS trainees, suggesting that these tokens (either because of the familiarity with the training talkers or the nature of the word stimuli) may have posed less of a difficulty for these learners, in line with the results observed for the CG.

### 3.3.2 Generalization to real words

Since training made use of non-words only, perception of real words was assessed at pre-test and at post-test in order to have a measure for real word identification comparable to that of nonsense word identification. Correct identification percentages for L2 vowels embedded in real words at pre-test and post-test, and the corresponding gain scores, were calculated for each group. Statistical analyses were carried out on the increase in percentage points from pretest to posttest obtained by each group, as previously done for the nonsense words. The results are given in Table 7.

| | CONTROL | | DIS | | ID | |
|---|---|---|---|---|---|---|
| Real words | % | *SD* | % | *SD* | % | *SD* |
| PRE | 72.2 | 11 | 78.2 | 9.7 | 73.1 | 11.2 |
| POST | 79.5 | 10.3 | 79.7 | 11.1 | 88.5 | 9.5 |
| GAIN (increase in percentage points) | 7.3 | 9.2 | 1.5 | 11.7 | 15.4 | 8.8 |

Table 7. Percent correct identification in real words at pre-test and post-test per group (generalization to real words).

Interestingly, vowel identification scores were higher in real words than in nonsense words already at pretest (72-78% *vs*. 54-56% for non-words), indicating a close relationship between lexical and phonetic categories, as discussed in the last section. Despite the high scores at pretest, improvement from pre-test to post-test was still observed. The ID, the group that improved the most with non-words (26 percentage points increase), was also the group that obtained the greatest gains with real words (15 percentage points). The DIS training regime, on the other hand, did not seem to enhance the ability to identify sounds in real words, as DIS trainees only improved by 1.5 percentage points with the training received. This slight improvement is possibly connected to the fact that their scores were higher at pre-test (78.2%), indicating that there was less room for improvement. In the case of the controls, the learners seemed to improve more in real word identification than when identifying non-words (7.3% vs. 3.7%). The GLMM analysis on gain scores yielded a significant effect of group ($F(2,51)=8.953$, $p<.01$). Sequential Bonferroni pairwise comparisons confirmed that only the identification group outperformed the control group, $p<.05$. Moreover, the ID group outperformed the DIS, indicating that generalization to real words for the trained sounds only occurred after receiving identification training ($p<.01$).

The identification scores for each individual vowel in the real word condition by each group are presented in Table 8. The results show that the control group appeared to improve by more than the DIS group for three out of the five sounds, namely /æ/, /ɜː/ and /ʌ/. This is probably explained by the higher scores obtained by the DIS at the onset of the study. At post-test, however, the results for these vowels do not seem to differ much across the two groups. The ID, however, obtained numerically higher identification scores than the controls and the DIS group in the identification of /æ/, /ʌ/ and /ɜ:/ after training, and both experimental groups improved numerically more than the controls for the sound /iː/, although all three groups reached similar identification scores with both /iː/ and /ɪ/ at post-test. The pattern of difficulty on the whole matches the one found for non-word identification previously. The vowel /æ/ obtained the lowest scores, while /iː/, /ʌ/ and particularly /ɪ/ were more accurately perceived. Moreover, overall scores were higher with real word identification than with non-words, in particular regarding the sound /ɜ:/.

| | CONTROL | | DIS | | ID | |
|---|---|---|---|---|---|---|
| SOUND | PRE | POST | PRE | POST | PRE | POST |
| /æ/ | 50.8 | 57.0 | 63.2 | 63.9 | 47.5 | 83.1 |
| | (30.8) | (34.4) | (24.9) | (27.7) | (30.2) | (25.1) |
| /ʌ/ | 71.1 | 81.2 | 86.8 | 79.8 | 83.1 | 92.5 |
| | (30.8) | (20.9) | (24.0) | (23.9) | (9.6) | (9.5) |
| /iː/ | 82.0 | 79.7 | 72.9 | 78.4 | 73.1 | 77.5 |
| | (16.4) | (17.6) | (18.3) | (14.1) | (17.8) | (17.5) |
| /ɪ/ | 89.0 | 93.7 | 88.9 | 95.1 | 90 | 94.3 |
| | (22.5) | (22.1) | (13.5) | (8.7) | (9.6) | (9.5) |
| /ɜː/ | 67.9 | 85.9 | 79.2 | 81.2 | 71.9 | 95 |
| | (27) | (21.8) | (23.5) | (27.9) | (25.9) | (10.2) |

Table 8. Percent correct identification at pretest and posttest for each individual vowel per group (real words; standard deviations are given in parentheses).

### 3.4 Retention effects

Two months after the post-test, a delayed post-test (or retention test) was administered. The aim of this test was to assess the long-term effects of training. Given that fewer participants took part in this last phase of the study, the analyses only include the results of the trainees that completed all three tests (pretest, posttest, delayed test). This explains the difference in absolute values between the results reported here and in previous sections. The total number of participants at this phase was less homogeneous among groups, as there were 9 controls, 17 ID trainees and 12 DIS trainees. In the same fashion as in the analysis of generalization results, it was considered that retention had taken place when the delayed test results were greater than the pre-test results and did not differ from (or were greater than) the post-test results. All three groups obtained numerically similar scores at post-test and retention test (see Table 9). GLMM analyses with time as the fixed effect (pre-test, post-test and delayed post-test) for each group showed that there was no significant effect of time for CG ($F(2, 72)=1.84$, $p>.05$), confirming that this group performed similarly across all three testing times. Regarding the trained groups, the models in each case yielded a significant effect of time (ID: $F(2, 48)=51.35$, $p<.001$; DIS: $F(2, 33)=7.62$, $p<.01$) and Bonferroni adjusted pairwise comparisons confirmed that the performance at pre-test significantly differed from the performance at post-

test and delayed post-test ($p<.001$ in both cases). Importantly, the delayed post-test results did not differ from the post-test results, confirming that learning was retained for a period of two months for both groups.

| | CONTROL | | DIS | | ID | |
|---|---|---|---|---|---|---|
| Test | % | *SD* | % | *SD* | % | *SD* |
| PRE | 56.7 | 11.3 | 53.0 | 4.2 | 51.8 | 9.7 |
| POST | 61.9 | 11.1 | 62.8 | 9.4 | 79.7 | 9.3 |
| DELAYED POST | 63.3 | 14.0 | 60.4 | 8.2 | 80.1 | 8.3 |

Table 9. Percent correct identification at pre-test, post-test and delayed post-test per group (data from participants who completed all three tests).

## 4. Discussion

The goal of this study was to evaluate the efficiency of two types of perceptual tasks for improving L2 speakers' ability to identify and produce target L2 vowels. The results show that HVPT positively affected the perception of L2 vowels by Spanish/Catalan L2 learners of English, and this improvement was facilitated by both methods tested, answering the first research question of the study. The ID group improved by 26.3 percentage points from pre to post-test and the categorical DIS group improved by 9.8. The amount of gain for the two experimental groups was similar to (DIS) or greater than (ID) the range of improvement usually reported in the phonetic training literature, that is, around 10%-15% (Jamieson & Morosan, 1986; Flege, 1989; Logan & Pruitt, 1995; Flege, 1995b; Iverson & Evans, 2009; Shinohara & Iverson, 2018). In addition, instances of generalization and retention of learning were found with both methods, as discussed below.

Globally, the findings of the present study provide further evidence that HVPT is effective and that both ID and DIS tasks can make a contribution to L2 learning (Iverson et al., 2012; Shinohara & Iverson, 2018). Further, the results suggest that categorical DIS tasks can be effective for training L2 vowel perception, even if to a lesser extent than ID tasks. The findings challenge previous views on the lower efficacy of discrimination tasks that were solely based on auditory discrimination tasks (Strange & Dittmann, 1984), and are more in agreement with training studies that reported that ID and categorical DIS were equally efficient for improving the perception of English final stops (Flege, 1995b), the perception of English initial stops (Carlet, 2017), the perception of coda nasals (Nozawa, 2015), the

perception and production of the English /r/-/l/ contrast (Shinohara & Iverson, 2018), and the perception of Thai tones (Wayland & Li, 2008). Thus, the current study supports previous findings about the efficacy of a categorical DIS task and extends them to vowel perception. The positive effect of categorical DIS tasks may be related to the fact that, contrary to the auditory DIS task, the categorical DIS task exposes learners to a greater range of acoustic variability, which in turn may promote L2 categorization (Polka, 1992).

Nevertheless, the greater gains obtained for the ID group suggest a potential superiority of ID over categorical DIS for training L2 vowel perception. This result is in line with the findings of the only previous study comparing ID and DIS tasks for training L2 vowel perception (Nozawa, 2015). It is possible that a task familiarity effect may have played a role in the better performance for the ID trainees. Recall that at pretest and posttest perception was tested by means of an identification task only, potentially creating an advantage for the ID trainees. This is a limitation of the current study, as discussed below. Nonetheless, the large and significant difference between ID and DIS may not be only the result of familiarity with the task. A possible explanation for this advantage may lie in the fact ID tasks may promote between-category sensitivity and thus be more efficient for category identification, as opposed to ID tasks, which may enhance within-category sensitivity (Jamieson & Morosan, 1986; Logan & Pruitt, 1995). Moreover, ID and DIS may also differ in that DIS tasks may tap into lower levels of phonological encoding that may not contribute greatly to category formation, whereas identification may involve the type of phonological encoding that is crucial for L2 categorization (Iverson et al., 2003; Iverson et al., 2008, Iverson et al., 2012).

Another possible explanation for the superiority of the ID over the DIS training method for L2 vowel perception might be connected to the presence of labels in the ID task, i.e. the response alternatives. The presence of labels may have provided learners with the chance to focus on phonetic form (i.e., phonetic symbols and/or orthography), which has been reported to impact speech perception (Saito, 2015). Note that while identification is a covert task, in which the single category presented in each trial is directly compared with a pre-existing memory representation, discrimination is an overt process, where the two items to be compared are physically present (Bohn, 2002). Thus, the nature of the task implies that the feedback provided was also different. ID feedback provided precise information about the category that the stimulus belonged to. By

contrast, the feedback provided to DIS trainees simply informed them about whether or not the two stimuli previously heard belonged to the same category. DIS trainees were not explicitly told which category each sound belonged to. Furthermore, alongside the phonetic symbol, each label or response alternative in the ID training task also contained two keywords exemplifying the target sounds (e.g., /iː/ - cheese/leaf; /ɜː/- earth/ first). Thus, during each trial, the identification group was forced to relate the sound they heard to a given phonetic symbol and a familiar spelling and word. There may be a link between the use of phonetic symbols and orthographic representations and the generalization to real words.

As pointed out above, one limitation of the current study is the lack of a discrimination test in addition to the identification test, which means that only the ID trainees may have benefitted from a task familiarity effect. However, Flege (1995b) and Carlet (2017) also compared ID and DIS training and evaluated only identification and reported that DIS trainees did not differ significantly from ID trainees in the identification of English stop consonants, showing no task familiarity effect. Further, no task familiarity effect was evident in the results obtained by Nozawa (2015) on the identification of final nasals, as ID and DIS trainees obtained comparable results. Furthermore, a later training study using the same stimuli as the current study (Cebrian, Carlet, Gavaldà & Gorba, 2017) tested vowel trainees in both abilities (discrimination and identification), and revealed that ID enhanced the identification of vowel sounds to a greater extent than the DIS method did, extending the findings of the current study. Interestingly, the ID method enhanced learners' vowel discrimination abilities to a similar extent as the DIS method did, also in line with previous findings (Wayland & Li, 2008). Hence, the preliminary findings of Cebrian et al. (2017) confirm the superiority of the ID training method for L2 vowel identification, as this method was able to enhance both perceptual abilities (identification and discrimination) either to a similar or to a greater extent than the categorical DIS method did.

The second research question involved the effect of high variability perceptual training on L2 vowel production. The results showed that, although numerically not large, there was a significant improvement after only 5 short sessions (30-mins) of perceptual training. This result corroborates previous findings that perceptual training may alter the production of L2 sounds, at least to some extent, without the need of explicit production training (Bradlow et al., 1997; Flege, 1989; Lambacher et al, 2005; Lengeris, 2008; Iverson et al., 2012; Thomson, 2011; Pereira, 2014; Rato &

Rauber, 2015, Shinohara & Iverson, 2018). The production results add to the observed superiority of the ID method over the DIS method for training vowel sounds, as the improvement experienced by the DIS group reached marginal significance only. It is possible that the differences between ID and DIS tasks discussed above also account for the different results regarding production. An additional explanation could stem from the fact that since production assessment included real words, the orthographic representation present in the labels of the ID training might have played a role. Also, recall that ID training was the only method that promoted generalization to real word stimuli. Thus it may follow that the group that experienced an improvement in the perception of real words also showed evidence of gains in the production of real words. Taken together, these findings fit the predictions of the SLM and the NLM, which postulate that perception gains occur prior to production gains and the former is a prerequisite for the latter. However, it seems that the learners were at the stage where perception is more developed than production, since the perceptual gains were overall greater than the production gains in the study. Thus, this result provides further evidence that the improvements in both domains do not seem to occur in parallel (Bradlow et al., 1997; Pereira, 2014; Iverson et al., 2012; cf. Rochet, 1995; Shinohara & Iverson, 2018); that is, changes in perception and production seem to develop differently.

The third research question addressed the possible differences between ID and DIS regarding generalization and retention effects. According to Flege (1995b) "a high degree of generalization suggests that a training procedure has engendered the formation of a long-term memory representation that is more abstract than the sum total of the physical properties encountered in the training stimuli" (p. 435). In the case of generalization to novel non-word stimuli produced by familiar talkers, the gain obtained during training was maintained or even increased a week later by both groups, providing evidence of robustness of learning (Logan & Pruitt, 1995). This result emphasizes the reported benefits of HVPT (Logan et al., 1991; Iverson et al., 2012; Shinohara & Iverson, 2018; among many others) and adds to previous findings that attest that both training methods (ID and categorical DIS) are effective (Flege, 1995b). The outcome is different, however, when we consider generalization to real words. First, it is relevant to note that perception of real words was better than perception of non-words, already at pretest. This may indicate that learners found it easier to recognize the vowels when they were found in words that they recognized. This may be related to the interplay between

lexical and phonological categories. Solé (2013) found that L2 contrasts that are not easily distinguishable in non-words may be differentiated in real words, indicating that L2 phonological categories may be formed after lexical categories, which are learned as a whole. Secondly, the ID was the only group that outperformed the controls and, thus, the only group that generalized the learning acquired through training to real words. The DIS group's performance with real words at pre-test was numerically higher than the ID's (DIS: 78% *vs*. ID: 73%), and there was no change after training (DIS: 79% *vs*. ID: 86%). Methodological differences discussed above, such as the covert nature and presence of labels in the case of the ID task, may account for difference between ID and DIS with real word perception.

Finally, both ID and DIS training methods were found to promote retention of learning after a period of two months, in line with several previous studies showing long-term effects of training (Bradlow et al., 1997,1999; Lively et al., 1993; Wang, 2002; Wang & Munro, 2004; Nishi & Kewley-Port, 2007; Rato, 2014). According to Flege (1995b), if knowledge acquired during training is retained over time, it may indicate that robust L2 categories have been established in the L2 learners' perceptual space. Moreover, this effect adds to the potential of phonetic training as an L2 teaching tool. All in all, the results of the delayed post-test confirm that both training methods (ID and categorical AX DIS) were able to promote long term effects and are effective when training vowel perception, in line with Flege's (1995b) findings on the perception of final stops. Moreover, the effects were retained over time, which may be an indicator of L2 category formation (Flege, 1995b).

## 5. Conclusions and implications

This study assessed the effect of two perceptual training methods (identification and same/different categorical discrimination) on the ability to identify and produce L2 vowels. The results showed positive changes in L2 learners' perceptual and production abilities as a result of high variability phonetic training (HVPT). Specifically, the present study provided evidence that both methods are effective, as both groups of trainees outperformed a group of untrained controls in the identification of trained sounds produced by untrained talkers, and both groups showed evidence of generalization to new non-word stimuli and retention of learning. However, the current study also evidenced that identification training was more effective in promoting generalization to perception of real words and

in improving vowel production, as judged by native speaker raters. In line with these results, a combination of both tasks (ID and categorical DIS) is suggested in order to enhance different perceptual abilities and maximize the effects of training. In fact, it has been suggested that discrimination tasks could be more suitable early in the learning process when the basic dimensions of variability are being discovered (Logan & Pruitt, 1995). Moreover, Pisoni and Lively (1995) explain that both types of training can be used in order to improve different perceptual skills. While identification training improves an "acquired equivalence", discrimination training improves an "acquired distinctiveness" (p. 445). Shinohara and Iverson (2018) argue that although both ID and categorical DIS are effective as training methods, DIS training may be easier to implement with lower proficiency learners who may not have acquired different categories for L2 sounds yet and/or for young learners who may have trouble with the use of labels. However, other studies have provided evidence that ID is favoured over DIS by L2 learners since the latter is found to be harder and somewhat tedious (Flege, 1995b; Carlet, 2017). In brief, the results of this study show that HVPT can be an efficient tool to enhance learners' perception and production abilities and that both ID and DIS may contribute to the learning process. Unfortunately, despite the success of HVPT, phonetic training methods are rarely implemented in the classroom. There is a need to bridge HVPT research and teaching practices, by making sure that this powerful perceptual tool is pedagogically implemented.

**Acknowledgments**

This research was made possible by the PhD grant PIF (429-02-1/2011) to the first author, the research grants from the Spanish Ministry of Economy and Competitiveness (FFI2013-46354-P and FFI2017-88016-P) and by a grant from the Catalan Government (2017SGR34).

**References**

Aliaga-García, Cristina & Joan Carles Mora. (2009). Assessing the effects of phonetic training on L2 sound perception and production. *Recent research in second language phonetics/phonology: Perception and production*, 2-31. Newcastle upon Tyne: Cambridge Scholars Publishing.

Aliaga-García, Cristina, Joan Carles Mora & Eva Cerviño-Povedano. (2011). L2 speech learning in adulthood and phonological short-term memory. *Poznań Studies in Contemporary Linguistics, 47*, 1-14.

Best, Catherine & Tyler, Michael. (2007). Non-native and second-language speech

perception: Commonalities and complementarities. In Ocke-Schwen Bohn & Murray J. Munro (eds.), *Language Experience in Second Language Speech Learning,* 13-34. Amsterdam/Philadelphia: John Benjamins.

Bohn, Ocke-Schwen. (2002). On phonetic similarity. In Petra Burmeister, Thorsten Piske & Andreas Rohde (eds.). *An Integrated View of Language Development: Papers in Honor of Henning Wode*, 191-216. Trier: Wissenschaftlicher Verlag.

Bohn, Ocke-Schwen & Munro, Murray J. (2007). *Language Experience in Second Language Speech Learning*. Amsterdam/Philadelphia: John Benjamins.

Bradlow, Ann R. (2008). Training non-native language sound patterns: Lessons from training Japanese adults on the English /r/ - /l/ contrast. In J. G. Hansen Edwards, & M. L. Zampini (eds.), *Phonology* and *Second Language Acquisition*, 287-308. Philadelphia: John Benjamins.

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

Carlet, Angelica. (2017). L2 perception and production of English consonants and vowels by Catalan speakers: The effects of attention and training task in a cross-training study". Unpublished PhD dissertation. Universitat Autònoma de Barcelona.

Cebrian, Juli. (2006). Experience and the use of non-native duration in L2 vowel categorization. *Journal of Phonetics 34*, 372-387.

Cebrian, Juli & Angelica Carlet. (2014). Second language learners' identification of target language phonemes: A short-term phonetic training study. *Canadian Modern Language Review 70*(4), 474-499.

Cebrian, Juli, Angelica Carlet, Núria Gavaldà & Celia Gorba. (2017). L2 vowel learning through perceptual training: Assessing training method, task familiarity and metalinguistic knowledge. Paper presented at the 18th World Congress of Applied Linguistics, Rio de Janeiro, Brazil.

Cebrian, Juli, Joan Carles Mora & Cristina Aliaga-García. (2011). Assessing crosslinguistic similarity by means of rated discrimination and perceptual assimilation tasks. In Magdalena Wrembel, Malgorzata Kul & Katarzyna Dziubalska-Kolaczyk (eds.), *Achievements and perspectives in the acquisition of second language speech: New Sounds 2010*, Volume I, 41-52. Frankfurt am Main: Peter Lang.

Cooke, Martin, María Luisa García-Lecumberri, John Maidment & Anders Ericsson. (2005). The web transcription tool. Retrieved February 18, 2017, from http://www.wtt.org.uk/.

Council of Europe (2001). *The common European framework of reference for languages: learning, teaching, assessment*. Cambridge, UK: Cambridge University Press.

Earle, F. Sayako & Emily B. Myers. (2014). Building phonetic categories: an

argument for the role of sleep. *Frontiers in psychology*, *5*, 1192-1192.
Flege, James E. (1989). Chinese subjects‟ perception of the word - final English /t/–/d/ contrast: Performance before and after training. *Journal of the Acoustical Society of America*, *86*(5), 1684-1697.
Flege, James E. (1991). Age of learning affects the authenticity of voice onset time (VOT) in stop consonants produced in a second language. *Journal of the Acoustical Society of America*, *89*(1), 395-411.
Flege, James E. (1995a). Second language speech learning: Theory, findings and problems. In Strange (ed.), 233-277.
Flege, James E. (1995b). Two procedures for training a novel second language phonetic contrast. *Applied Psycholinguistics*, *16*, 425-442.
Flege, James E., Ocke-Schwen Bohn & Sunyoung Jang. (1997). Effects of experience on non-native speakers‟ production and perception of English vowels. *Journal of Phonetics, 25*(4), 437-470.
Flege, James E. (2003). Assessing constraints on L2 segmental production and perception. In A. Meyer and N. Schiller (eds.). *Phonetics and Phonology in Language Comprehension and Production, Differences and Similarities*, 320-355. Berlin: Mouton de Gruyter.
Grosjean, François. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics, 28*(4), 267-283.
Hardison, Debra M. (2004). Generalization of computer-assisted prosody training: Quantitative and qualitative findings. *Language Learning & Technology*, *8*(1), 34-52.
Højen, Anders & James E. Flege. (2006). Early learners' discrimination of second-language vowels. *The Journal of the Acoustical Society of America, 119*(5), 3072-3084.
Ingram, John C. & See-Gyoon Park. (1997). Cross-language vowel perception and production by Japanese and Korean learners of English. *Journal of Phonetics, 25*(3), 343-370.
Iverson, Paul & Bronwen G. Evans. (2007). Learning English vowels with different first-language vowel systems: Perception of formant targets, formant movement, and duration. *The Journal of the Acoustical Society of America, 122*(5), 2842-2854.
Iverson, Paul & Bronwen G. Evans. (2009). Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers. *The Journal of the Acoustical Society of America, 126*(2), 866-877.
Iverson, Paul, Melanie Pinet & Bronwen G. Evans. (2012). Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics, 33*(01), 145-160.
Jamieson, Donald G. & David E. Morosan. (1986). Training non-native speech contrasts in adults: Acquisition of the English /ð/-/θ/ contrast by francophones. *Perception & Psychophysics, 40*(4), 205-215.
Kuhl, Patricia K. & Paul Iverson. (1995). Linguistic experience and the perceptual

magnet effect. In Strange (ed.), 121-154.

Lacabex, Esther G., María Luisa García-Lecumberri & Martin Cooke. (2008). Identification of the contrast full vowel-schwa: training effects and generalization to a new perceptual context. *Ilha do Desterro A Journal of English Language, Literatures in English and Cultural Studies, 55*, 173-196.

Lambacher, Stephen G., William L. Martens, Kazuhiko Kakehi, Chandrajith A. Marasinghe & Garry Molholt. (2005). The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics, 26*(02), 227-247.

Lengeris, Angelos. (2008). The effectiveness of auditory phonetic training on Greek native speakers' perception and production of southern British English vowels. In *Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics, ExLing 2008,* 133-136. Athens: ISCA.

Lively, Scot E., John S. Logan & David B. Pisoni. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America, 94*(3), 1242-1255.

Logan, John S., Scott E. Lively & David B. Pisoni. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America, 89*(2), 874-886.

Logan, John S. & John S. Pruitt. (1995). Methodological issues in training listeners to perceive non- native phonemes. In Strange (ed.), 351-378.

McClaskey, Cynthia L., David B. Pisoni & Thomas D. Carrell. (1983). Transfer of training of a new linguistic contrast in voicing. *Perception & Psychophysics, 34*(4), 323-330.

Muñoz, Carmen. (2008). Symmetries and asymmetries of age effects in naturalistic and instructed L2 learning. *Applied Linguistics*, *29*(4), 578-596.

Munro, Murray J. (2008). Foreign accent and speech intelligibility. In Jette G. Hansen Edwards, & Mary L. Zampini (eds.), *Phonology* and *Second Language Acquisition*, 193-218. Philadelphia: John Benjamins.

Nishi, Kanae & Diane Kewley-Port. (2007). Training Japanese listeners to perceive American English vowels: Influence of training sets. *Journal of Speech Language and Hearing Research*, *50*(6), 1496-1509.

Nobre-Oliveira, Denize. (2007). Effects of perceptual training on the learning of English vowels in non-native settings. In *Proceedings of the 5th International symposium on the acquisition of second language speech, New Sounds,* Vol. 5, 382-389. Florianopolis: Federal University of Santa Catarina.

Nozawa, Takeshi. (2015). Effects of training methods and attention on the identification and discrimination of American English coda nasals by native Japanese listeners. In Scottish Consortium for ICPhS 2015 (ed.), *Proceedings of the 18th International Congress of Phonetic Sciences.* Glasgow, UK: the University of Glasgow.

Pereira, Yasna I. (2014). *Perception and production of English vowels by Chilean*

*learners of English: Effect of auditory and visual modalities on phonetic training* (Unpublished doctoral dissertation). University College London, London, UK.

Piske, Thorsten, Ian R.A. MacKay & James E. Flege. (2001). Factors affecting degree of foreign accent in an L2: a review. *Journal of Phonetics*, *29*, 191-215.

Pisoni, David B. & Scott E. Lively. (1995). Variability and invariance in speech perception: A new look at some old problems in perceptual learning. In Strange (ed.), 433-462.

Pisoni, David B., Richard N. Aslin, Alan J. Perey & Beth L. Hennessy. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception and Performance, 8*(2), 297-314.

Polka, Linda. (1992). Characterizing the influence of native language experience on adult speech perception. *Perception & Psychophysics, 52*(1), 37-52.

Rato, Anabela. (2014). Effects of Perceptual Training on the Identification of English Vowels by Native Speakers of European Portuguese. *Concordia Working Papers in Applied Linguistics, 5*, 529-546.

Rato, Anabela & Andreia Rauber. (2015). The effects of perceptual training on the production of English vowel contrasts by Portuguese learners. In The Scottish Consortium for ICPhS 2015 (ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: the University of Glasgow.

Rauber, Andreia, Anabela Rato, Denise Kluge & Giane Santos. (2012). TP (Version 3.1).[Software]. *Brazil: Worken.* [http://www.worken.com.br/tp_regfree.php?l=i].

Rochet, Bernard L. (1995). Perception and production of L2 speech sounds by adults. In Strange (ed.), 379-410.

Saito, Kazuya. (2015). Variables affecting the effects of recasts on L2 pronunciation development. *Language Teaching Research*, *19*(3), 276-300.

Shinohara, Yasuaki & Paul Iverson. (2018). High variability identification and discrimination training for Japanese speakers learning English /r/–/l/. *Journal of Phonetics 66*, 242-251.

Solé, Maria Josep. (2013). Phonological vs. lexical categories in an L2. In *Proceedings of the 6th Phonetics and Phonology in Iberia Conference,* 58-59. Lisbon, Portugal, Lisbon University.

Strange, Winifred (ed.). (1995). *Speech Perception and Linguistic Experience: Issues in Cross Language Research*. Timonium, MD: York Press.

Strange, Winifred & Sybilla Dittmann. (1984). Effects of discrimination training on the perception of /r- l/ by Japanese adults learning English. *Perception & Psychophysics, 36*(2), 131-145.

Thomson, Ron I. (2012). Improving L2 listeners' perception of English vowels: A computer- mediated approach. *Language Learning, 62*(4), 1231-1258.

Wang, Xinchun & Murray J. Munro. (2004). Computer-based training for learning English vowel contrasts. *System, 32*(4), 539-552.

Wayland, Ratree P. & Bin Li. (2008). Effects of two training procedures in cross-language perception of tones. *Journal of Phonetics, 36*(2), 250-267.

**Appendix 1 –Perception and production stimuli**

| **Training stimuli** | | | | | |
|---|---|---|---|---|---|
| **/æ-ʌ/** | | **/ɪ-i/** | | **/ɜː-e/, /ɜː-ɑː/** | |
| dadge | tadge | deege | teege | darge | targe |
| dudge | tudge | didge | tidge | derge | terge |
| pav | bav | peedge | beedge | parsh | barsh |
| puv | buv | pidge | bidge | persh | bersh |
| kak | gak | keedge | geedge | karch | garch |
| kuk | guk | kidge | gidge | kerch | gerch |
| zat | zad | jeet | jeed | zart | zard |
| zut | zud | jit | jid | zert | zerd |
| vap | vab | veep | veeb | jarp | jarb |
| vup | vub | vip | vib | jerp | jerb |
| vak | vag | veek | veeg | vark | varg |
| vuk | vug | vik | vig | verk | verg |
| **Testing stimuli** | | | | | |
| **/æ-ʌ/** | | **/ɪ-i/** | | **/ɜː/** | |
| vab | vap | veeb | veep | jurb | |
| zad | zat | jeed | jeet | jerd | |
| vag | vack | veeg | veek | verg | |
| vub | vup | vib | vip | jurp | |
| zud | zut | jid | jit | jurt | |
| vugg | vuck | vig | vick | verk | |
| **Generalization to real words stimuli** | | | | | |
| **/æ-ʌ/** | | **/ɪ-i/** | | **/ɜː/** | |
| cap | cab | feet | feed | hurt | heard |
| pup | pub | bit | bid | | |
| **Generalization to novel non-words stimuli** | | | | | |
| **/æ-ʌ/** | | **/ɪ-i/** | | **/ɜː/** | |
| dack | pag | fip | pid | vert | derg |
| dut | Jud | geep | keeb | | |
| **Production elicitation list** | | | | | |
| **/æ-ʌ/** | | **/ɪ-i/** | | **/ɜː/** | |
| cap | cab | bit | bid | hurt | heard |
| buck | bug | feet | feed | | |