Language Learning & Technology
ISSN 1094-3501

February 2020, Volume 24, Issue 1
pp. 62–85

ARTICLE

Using apps for pronunciation training: An empirical evaluation of the English File Pronunciation app

Jonás Fouz-González, *University of Murcia*

Abstract

This study explores the potential of the English File Pronunciation (EFP) app to help foreign language learners improve their pronunciation. Participants were 52 Spanish EFL learners enrolled in an English Studies degree. Pre- and post-tests were used to assess the participants' perception and production (imitative, controlled, and spontaneous) before and after training. The targets addressed were a range of segmental features that tend to be fossilised in the interlanguage of advanced Spanish EFL learners, namely English /æ ɑː ʌ ə/ *and the* /s – z/ *contrast. Training took place over a period of two weeks in which participants used the English File pronunciation app for around 20 minutes a day. Participants were randomly assigned to two groups (control and experimental). However, after the post-test, the group that had acted as control started to receive instruction and, after two weeks, took a second post-test, therefore acting as experimental too. Training fostered substantial improvements in the learners' perception and production of the target features, although the differences between groups were not statistically significant for every sound or in every task.*

Keywords: *Pronunciation, Second Language Acquisition (SLA), Computer Assisted Pronunciation Training (CAPT), Mobile Assisted Language Learning (MALL)*

Language(s) Learned in This Study: *English*

APA Citation: Fouz-González, J. (2020). Using apps for pronunciation training: An empirical evaluation of the English File Pronunciation app. *Language Learning & Technology, 24*(1), 62–85.
https://doi.org/10125/44709

Introduction

Mastering the pronunciation of a second (L2) or foreign language (FL) is an extremely challenging task for learners, given that success does not only depend on the learners' effort or declarative knowledge, but on an interplay of perceptual, psychomotor, cognitive, and affective factors (Pennington, 1998). One of the biggest obstacles learners face is perceiving the FL phonology adequately, as their perception is strongly conditioned by the phonological system of their L1 (Best & Tyler, 2007; Flege, 1995). Moreover, if learners do not have "accurate perceptual 'targets' to guide the sensorimotor learning of L2 sounds, productions of the L2 sounds will be inaccurate", as they will resort to the same articulatory movements that they use for the articulation of L1 sounds (Flege, 1995, p. 238).

Under the assumption that an adequate perception of the pronunciation of the FL plays a key role in subsequent accurate productions, numerous researchers have addressed the potential of perceptual training to help learners perceive and produce aspects of the FL pronunciation that are difficult to master without instruction. The results have been generally positive, showing that perceptual training paradigms can help learners improve their perception (Gómez-Lacabex, García-Lecumberri, & Cooke, 2008; Logan, Lively, & Pisoni, 1991) and production (Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Carlet, 2017; Rato, 2013; Thomson, 2011) of the target features, even when production is not trained.

Given the challenging nature of pronunciation, the limited availability of authentic input in FL contexts, and the little time devoted to pronunciation in EFL classes, technology has become a strong ally for pronunciation work. Computer-Assisted Pronunciation Training (CAPT) research has shown the numerous possibilities different tools offer to help learners improve their perception and production of different segmental and suprasegmental features, either by enhancing their perception of the features themselves, or by highlighting relevant aspects of the learners' output to help them notice their mistakes and how to pronounce the targets (see Fouz-González, 2015 for a review). As a case in point, studies have shown the potential of spectrograms (Olson, 2014) and waveform displays (Motohashi-Saigo & Hardison, 2009) to help learners notice relevant aspects of pronunciation and promote improvements in the learners' production of the target features. Pitch contours have also been useful in raising the learners' awareness of the prosodic organisation of speech and of the communicative function of intonation (Ramírez-Verdugo, 2006), enhancing the acquisition of prosody and fostering generalisations to segmental accuracy and to novel sentences (Hardison, 2004). Additionally, studies have also explored the potential of ASR feedback and, even though the gains fostered by the ASR feedback have not always been found to be significantly different from those obtained through other types of feedback, the findings suggest that ASR-based training can help learners work on their pronunciation of challenging segmental features (Neri, Cucchiarini, & Strik, 2008).

Notwithstanding the above, and despite the enormous potential these tools hold for certain purposes and contexts, some of them are not entirely suitable for every learner or for autonomous practice, given the difficulty in interpreting the feedback offered without specific training (e.g. in the case of spectrograms or waveforms), the lack of clear indications on how to improve when mistakes are detected by tools using automatic error detection, or the impossibility to provide accurate feedback on spontaneous speech (Levis, 2007; O'Brien, 2011; Pennington, 1999).

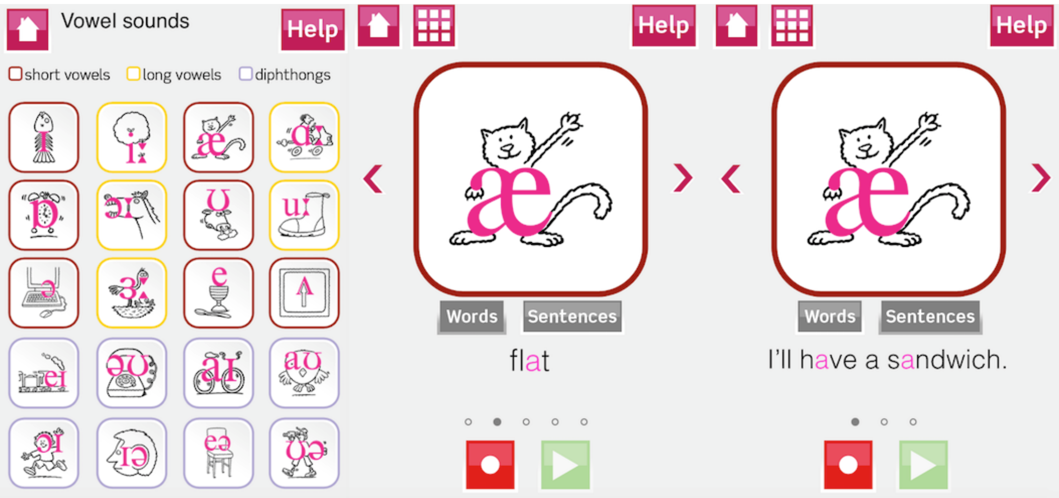
Because of the need to control for as many variables as possible in order to ensure the reliability of the studies conducted, Computer Assisted Language Learning (CALL) and CAPT studies have often been conducted in rather controlled, laboratory-like environments. Nevertheless, one of the main advantages of CALL, Mobile Assisted Language Learning (MALL) and, by extension CAPT, is the fact that learners can practise at their own pace, at a time and location of their choosing. Hence, one way of bringing pronunciation training to the learners' fingertips is through the use of their own mobile devices.

Focusing on pronunciation, research has demonstrated the potential of mobile-based High Variability Phonetic Training (HVPT) to help learners improve their perception of challenging sound contrasts (Uther, Uther, Athanasopoulos, Singh, & Akahane-Yamada, 2007), of mobile speech recognisers to provide learners' with feedback on controlled production (Liakin, Cardoso, & Liakina, 2015), or of shadowing practice using iPods to work on the learners' comprehensibility and fluency (Foote & McDonough, 2017). Because pronunciation is such a challenging competence for FL learners, so often neglected in FL classrooms, and given that one of the most-cited advantages of technology is that learners can practise anytime, anywhere, research should continue to explore the learning potential of different tools and techniques when learners use them outside the classroom. Given this, and in an attempt to explore tools that are easily accessible and easy to use for every learner, this study was set up to explore the potential of the English File Pronunciation app (henceforth EFP app; Oxford University Press, 2012) to help FL learners improve their perception and production of a range of English sounds.

The Present Study

As Colpaert (2004) notes, hype is often achieved in CALL when amateurs, not trained professionals, are able to develop their own applications. Training paradigms like the ones by Uther et al. (2007), Qian, Chukharev-Hudilainen, and Levis (2018), or Thomson's (2018) web-based application are exemplary, since they have been designed by pronunciation experts and are grounded in research. However, because it is not always possible for teachers to design their own applications, research should also investigate the possibilities and the actual learning potential commercial apps offer.

The EFP app includes an interactive sound chart illustrating the sounds of English and two activities. The sound chart (Figure 1, left) uses the same phonetic symbols with pictorial illustrations offered in the *English File* collection of books. Users can hear the sound in sample words (Figure 1, middle) and sentences (Figure 1, right), with the spelling featuring the target sound in a different colour. Additionally, users are also offered the opportunity to record themselves and compare their recording to the model.

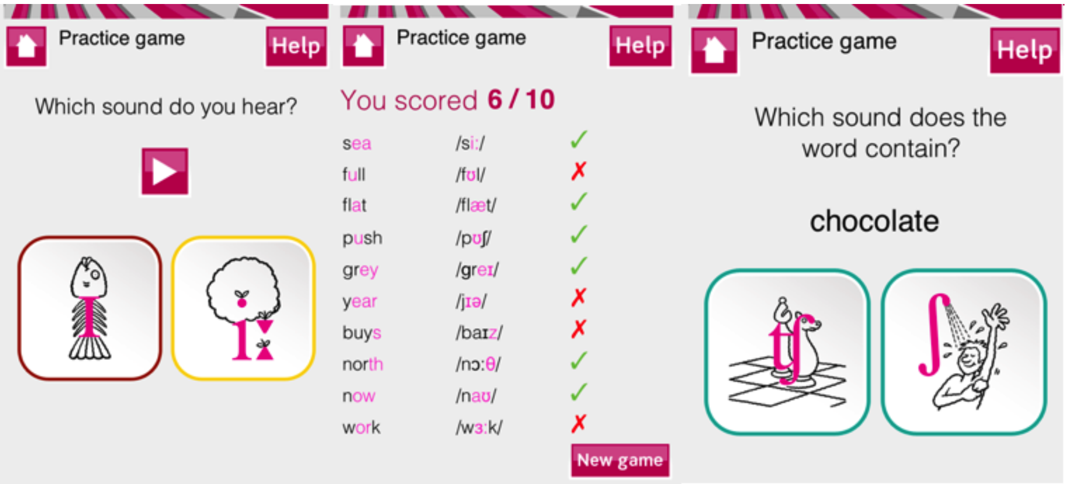


flat

I'll have a sandwich.

Figure 1. Screenshots from the EFP app: Sound chart (left), sample word (middle) and sample sentence (right).

The first activity is a sound identification activity in which users listen to words in isolation (no orthographic representation is provided) and have to identify the sound they hear out of two possible options displayed as phonetic symbols (Figure 2, left). Users are offered immediate feedback on their responses (a green tick if their answer is correct, a red cross if their answer is wrong). Every 10 words, a progress screen shows the scores for those words and a summary of the user's responses (Figure 2, middle). In the second activity, users are presented with words in isolation and have to decide which of the two sounds that appear on the screen is featured in the target words; users are shown the word in orthography, but they cannot hear the word (Figure 2, right).



Which sound do you hear?

You scored **6 / 10**

sea	/siː/	✓
full	/fʊl/	✗
flat	/flæt/	✓
push	/pʊʃ/	✓
grey	/greɪ/	✓
year	/jɪə/	✗
buys	/baɪz/	✗
north	/nɔːθ/	✓
now	/naʊ/	✓
work	/wɜːk/	✗

Which sound does the word contain?

chocolate

Figure 2. Screenshots from the EFP app. Activity 1 (left), progress screen (middle) and activity 2 (right).

Research Questions

RQ1. Can instruction through the EFP app help learners improve their perception of the target sounds?

RQ2. Can instruction through the EFP app help learners improve their production of the target sounds?

Method

Participants

Participants were recruited from a phonetics course in an English studies degree program at the University of Murcia (Spain). They were 54 students, 41 females and 13 males (mean age = 19.3; *SD* = 0.6).[1] In the questionnaires administered at the beginning of the study, participants reported having a B2 level according to the Common European Framework of Reference for Languages (CEFR). They had completed a B2 course in the first year of their degree and they were now in the first of two courses preparing them for C1.

Target Features

The target items addressed were four English vowel sounds (/æ ɑː ʌ ə/) and the /s – z/ contrast. In a study investigating the interlanguage of a group of students with the same profile as the ones in this study, Monroy-Casas (2001) offered a comprehensive account of different segmental substitutions affecting the aforementioned targets. These include substitutions such as English /æ/ for Spanish /a/, as in *family* *[ˈfamili]; English /ə/ and /ʌ/ for Spanish /a/, as in *another* *[aˈnaðar]; or English /ɑː/ for Spanish /a/, as in *castle* *[ˈkasel]. In the case of /ə/, the influence of orthography causes a wide range of potential substitutions for different Spanish vowels, such as *[ˈmarβelus] for *marvelous* (/ˈmɑːvələs/), *[poˈlisman] for *policeman* (/pəˈliːsmən/), or *[teleˈβision] for *television* (Monroy-Casas, 2001). Regarding the English /s – z/ contrast, although /s/ may be realised phonetically as [z] due to assimilation processes, Castilian Spanish only has one alveolar fricative in its phonemic repertoire (Hualde, 2014), the voiceless /s/. Thus, Spanish EFL learners often fail to mark the distinction between the two, pronouncing words like *noises* /ˈnɔɪzɪz/ as *[ˈnoises]; *was* /wəz/ as *[wos]; or *girls* /gɜːlz/ as *[gels] (Monroy-Casas, 2001).

These aspects were selected because they tend to be fossilised in the interlanguage of advanced Spanish EFL learners and were known to be problematic for the target group. It is important to note that participants in this study had a B2 level according to the CEFR and were thus considered to be generally intelligible. The CEFR's recently redeveloped scale for phonological control defines B2-level students' overall phonological control as "[c]an generally use appropriate intonation, place stress correctly and articulate individual sounds clearly; accent tends to be influenced by other language(s) he/she speaks, but has little or no effect on intelligibility" (Council of Europe, 2018, p. 136). Therefore, although participants in this study should not have problems with intelligibility in general, it was considered important to help them improve their pronunciation of features they tend to mispronounce systematically, as the C1 level (the one for which they were preparing) implies that students "can articulate virtually all of the sounds of the target language with a high degree of control" (Council of Europe, 2018, p. 136). Hence, the goal of this study was not to assess the impact of training in the learners' intelligibility, comprehensibility or accentedness, but to measure the type of improvements the app could foster in the learners' perception and production of the target sounds.

Study Design

One of the challenges in studies investigating the potential of a given approach that is considered to be beneficial for students is to be able to use a control group without depriving participants of instruction (Lee, Jang, & Plonsky, 2015; Lord, 2008; Thomson & Derwing, 2015). A possible solution is to use both groups as control and experimental at the same time, with each group focusing on different aspects (Fouz-González, 2019). This has several advantages, such as the fact that no group is deprived of training, that the effectiveness of the approach can be measured with a larger sample, or that participants in both groups receive similar amounts of extra exposure, therefore making the focus of training the only difference between groups. Given that the app under examination in this study does not offer users the possibility to choose the sounds they want to practise and therefore all users are exposed to the same (full) training set, it was not possible to conduct training simultaneously and create different training conditions for each group. Hence, from pre- to post-test, group 1 (G1) acted as experimental and group 2 (G2) acted as control. However, once G2 had acted as control, they started to receive training too, therefore also acting as experimental (Figure 3).

Participants were randomly assigned to two groups (G1 = 27, G2 = 27). Participants in G1 were required to attend four meetings with the researcher, as perception and production were tested on different days at pre- and post-tests, and participants in G2 were asked to attend five. In order to avoid imposing excess demands on G2 as compared to G1, who finished the study earlier and had to attend a total of four tests, the second post-test for G2 only addressed perception.[2]

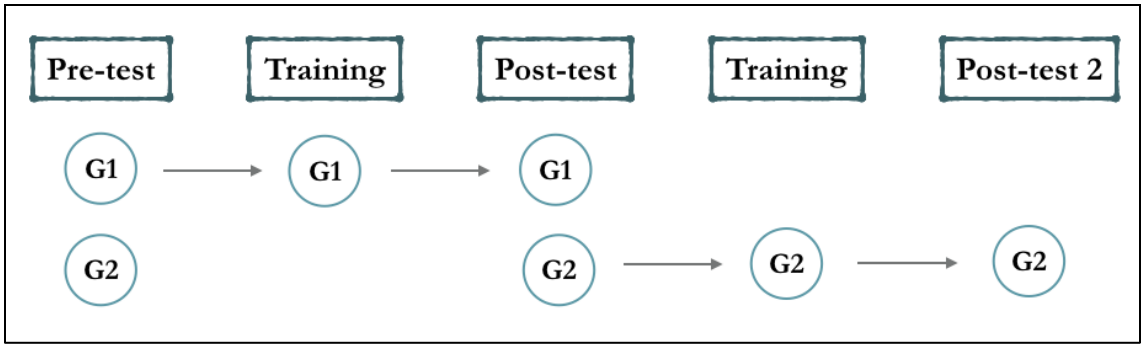


Figure 3. Study design.

Training Procedure

Training consisted in using the EFP app over a period of two weeks. Learners were allowed to use the app anywhere and at any time, but they were given some guidelines on how to use the app during the study to ensure that the amount of training participants received was similar. Participants were asked to complete 10 games a day (see Training stimuli section) on each of the two activities from Monday to Friday, which took approximately 15–20 minutes per day.

Participants were told what the target sounds were after the pre-test. They were asked to explore these sounds in the phonemic chart every day before completing the activities, then practise with activity 1, followed by activity 2. In order to control task completion, participants were asked to take screenshots of every progress screen and share them with the researcher through Dropbox. Screenshots show the time and date at which they were taken, which allowed the researcher to check that learners completed every activity on the day they were supposed to.

Training Stimuli

In the EFP app, users cannot control the input to which they are exposed during training. They are presented with sets of words featuring the whole range of sounds addressed in the activities. Thus, in order to estimate the amount of practice learners would have for each sound when using the app, the researcher conducted a trial run of the first 1000 stimuli for activity 1 and the first 500 for activity 2.

Given that progress screens appear every 10 stimuli, in order to quantify the amount of training learners received, in this article, every 10 words will be referred to as a “game” and every 10 games will be referred to as a “level” (i.e. 100 words).[3] Table 1 shows the number of instantiations and the percentage of occurrence of each target sound over a set of 10 levels for activity 1 (i.e. 1000 stimuli) and 5 levels for activity 2 (i.e. 500 stimuli) based on the above-mentioned trial run. The percentages in the table illustrate the number of times every target sound was featured every 100 words (i.e. the daily exposure learners received on each activity). The target sounds were featured with different spellings and in different positions in the word (initial, medial, and final, though not for every sound).

Table 1. *Number of Instantiations and Percentage of Occurrence of the Target Sounds in the EFP App*

	Activity 1		**Activity 2**	
	(n = 1000)	**%**	**(n = 500)**	**%**
/æ/	68	6.8	31	6.2
/ɑ/	34	3.4	19	3.8
/ʌ/	32	3.2	18	3.6
/ə/	88	8.8	52	10.4
/s/	38	3.8	21	4.2
/z/	33	3.3	19	3.8

Perception and Production Tests

The learners' perception was measured with an oddity discrimination task and an identification task. In the oddity discrimination task, stimuli were presented in blocks of three minimally paired words that either had the same phonological composition (i.e. "catch triads" – *cat-cat-cat*), or one of them differed in one sound and learners have to identify the one that was different (i.e. "change triads" – *cat-cat-cut*). Stimuli in each triad were always pronounced by three different speakers, male and female. In the identification task, learners were presented with one stimulus at a time and had to identify the sounds they heard among a range of options, including the target sounds and distractors (Figure 4 shows the identification task with illustrations from the English File phonemic chart).

The learners' production was evaluated with three tasks aimed at measuring their imitative, controlled, and spontaneous pronunciation of the target features, namely an imitation task, a sentence-reading task, and a timed picture-description task. To ensure that learners pronounced the target words in the spontaneous task, the pictures participants had to describe were presented with several words to guide their description (including target words and distractors).

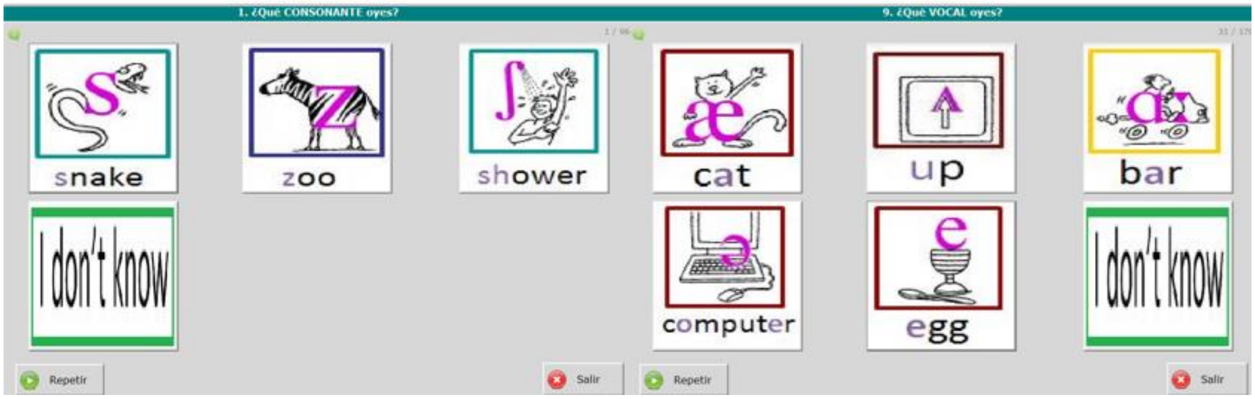


Figure 4. Sample screenshot of the identification task used for /s – z/ (left) and for /æ ʌ ɑː ə/ (right).

Testing Stimuli

Stimuli for the perception tests were obtained from several English dictionaries, with the exception of 13 items in the identification task featuring the /s – z/ contrast in plural words. Since audio illustrations in pronunciation dictionaries do not include plurals, these words were recorded by two speakers of standard British English, a female from Brighton (UK) and a male from Preston (UK).

Stimuli in the discrimination task consisted of 40 triads of minimally paired words and 10 triads with the strong and weak versions of words whose vowels can be reduced to schwa. The minimal pairs for vowel sounds were monosyllabic words with the target vowel as nucleus and surrounded by different voiced and voiceless consonants. Triads for schwa included the same word three times featuring its weak and strong versions (e.g. *that* [ðæt] vs. [ðət]). For the /s – z/ contrast, stimuli consisted of 15 minimally paired words, ten pairs were aimed at measuring the /s – z/ contrast and five were included as distractors contrasting /s – ʃ/ (Appendix A).

The discrimination test consisted of a total of 80 triads. There were 50 change triads (10 for each pair of targets and for schwa) and 25 catch triads (5 for each sound except for schwa, which could not be featured in five catch triads due to the impossibility of obtaining three different instantiations of the weak versions of words with schwa in the above-mentioned dictionaries). Five more triads were included as distractors featuring the /ʃ – s/ contrast.

Stimuli for the identification test consisted of 120 words and 5 distractors (Appendix B). Each target sound was featured in 20 words, 10 familiar and 10 novel. The criteria for the selection of familiar stimuli were the frequency of appearance of the words during the trial run explained above and the orthographic representations featuring the target sounds in those words. Novel stimuli featured the target sounds in different positions and with different spellings.

Regarding production, the testing stimuli featured five targets (/æ ɑː ʌ ə z/). The imitation task consisted of a total of 20 stimuli (4 per target sound) obtained from the same compilation used in the discrimination task featuring vowel sounds in different phonetic contexts. However, the stimuli for schwa were lexical items that are commonly mispronounced due to the influence of spelling. In the sentence-reading task, each sound was featured in 10 familiar words selected from the most commonly occurring words in the app.[4] Additionally, five novel words per sound were included in order to test possible generalisation gains. Finally, in the timed picture-description task, each target sound was assessed with four familiar tokens (Appendix C).[5]

Testing Procedure

Perception tests were conducted in a quiet computer room at the university using TP, an open-source application for developing and administering speech perception tasks (Rato, Rauber, Kluge, & Santos, 2015). During the test, learners were allowed to listen to each triad twice. The production tests were conducted individually in a quiet room at the university. They started with the sentence-reading task, followed by the timed picture-description task, and finally, the imitation task. The imitation task was administered last in order to avoid possible training effects for the other two tasks. The tests were recorded with a SAMSON C01U Microphone and a MacBook Pro computer. It is important to note that although the perception tests and their results are presented first in this article, the production tests were always conducted first in order to avoid possible training effects.

Evaluation of Production Data

The participants' productions were evaluated by three non-native judges expert in English pronunciation (L1 Spanish). A fourth judge was used in order to disambiguate disagreements. The judges were experienced EFL teachers. They held a five-year degree in English Philology and had taken graduate and undergraduate courses on English phonetics and phonology. Two of them held PhD degrees in English Linguistics (with a focus on phonetics) and the other two were completing their PhDs on the acquisition and learning of pronunciation.

The ratings were always dichotomous (1 if the target sound was pronounced adequately, 0 if it was mispronounced). Raters could play each stimulus as many times as they needed. Interrater reliability was assessed with Fleiss Kappa test. The test yielded a reliability measure of 0.954, which can be interpreted as "almost perfect agreement" (0.81-1.00 range). Intrarater reliability was assessed by analysing the judges' consistency in rating 20 extra items that had already been assessed (4 per target sound). There were no instances in which raters assigned a different score to an item that had already been assessed, so no extra tests were conducted.

Results

Perception

The pre- and post-test data were analysed with two-way mixed measures ANOVAs, with time as within-subjects factor and group as between-subjects factor. The data from the post-test and second post-test by G2 were analysed with paired t-tests, although for contrasts in which the data were not normally distributed, Wilcoxon-Signed-Rank tests were used. The analyses of the improvements made for each sound were done separately. No multiple comparisons were made in order to avoid losing statistical power.

For the sake of simplicity, when referring to G2's performance from pre- to post-test (i.e. when they acted as control), they will be referred to as G2C; when referring to their performance from post-test to the second post-test (i.e. when they acted as experimental), they will be referred to as G2E.[6] Standard deviations are always presented in brackets immediately after mean scores.

The analysis of the total scores in the discrimination task revealed a significant effect of time (*F*(1,47) = 15.79, *p* = <0.001, *r* = 0.5), but no interaction effects between time and group (*F*(1,47) = 0.199 , *p* = 0.65, *r* = 0.06); that is, improvements were made from pre- to post-test, but they were similar between groups. The analysis for G2E from post-test to second post-test shows that the differences did not reach significance either (*t* (22) = 0.83, *p* = 0.41, *d* = 0.17). Since no significant differences were found for the total scores in this task or in the scores for specific contrasts, the data for each contrast are not presented here.

As for the identification task, the results considering the total scores from pre- to post-test reveal a significant effect of time (*F*(1,44) = 91.03, *p* = <0.001, *r* = 0.82) and a significant interaction between time and group (*F*(1,44) = 25.36, *p* = <0.001, *r* = 0.6). The data show that the instruction had a positive effect on the learners' perception of the target features, which is further supported by the improvements made by G2E from the post-test to the second post-test (*t*(19) = 7.01 , *p* = < 0.001, *d* = 1.57). G2E's mean scores increased by 16.3 points (13.6%) after training, which is almost three times the improvement they made when acting as control (Figure 5).

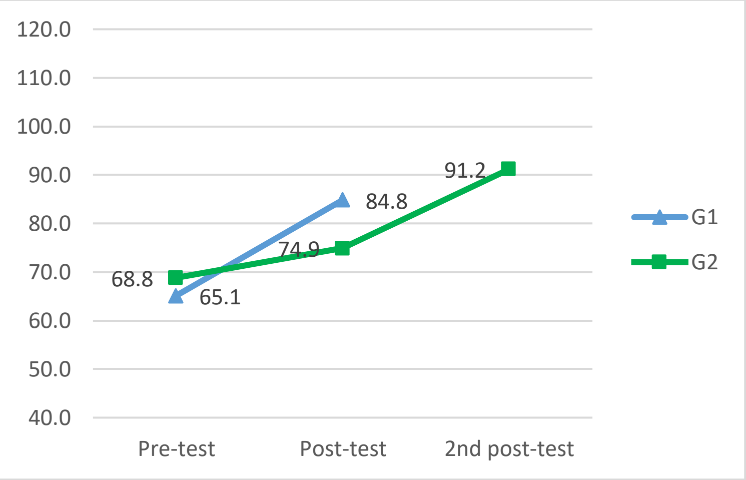


Figure 5. Mean scores in the identification task at pre-test, post-test and second post-test.

In order to explore the impact of instruction on the different target sounds, the scores obtained for each sound in the identification task for familiar and novel stimuli were analysed separately. The mean scores at pre-, post- and second post-tests for familiar and novel stimuli are illustrated visually in Figure 6 and Figure

7 respectively. Focusing on G1 and G2C's scores for familiar stimuli, the differences between groups were found to be significant for /ʌ/, /ɑː/, /ə/, and /z/, but not for /æ/ (Appendix D). Regarding G2E's scores, the differences between groups reached statistical significance for every target sound. As for novel stimuli, G1 and G2C's scores only revealed significant Time x Group interactions for /ʌ/. However, the scores by G2E revealed a significant effect for /ʌ/, /ə/ and /z/ (Appendix D). The mean scores obtained in the pre-test, post-test, and second post-test, the standard deviations and the 95% confidence intervals are presented in Appendix E.

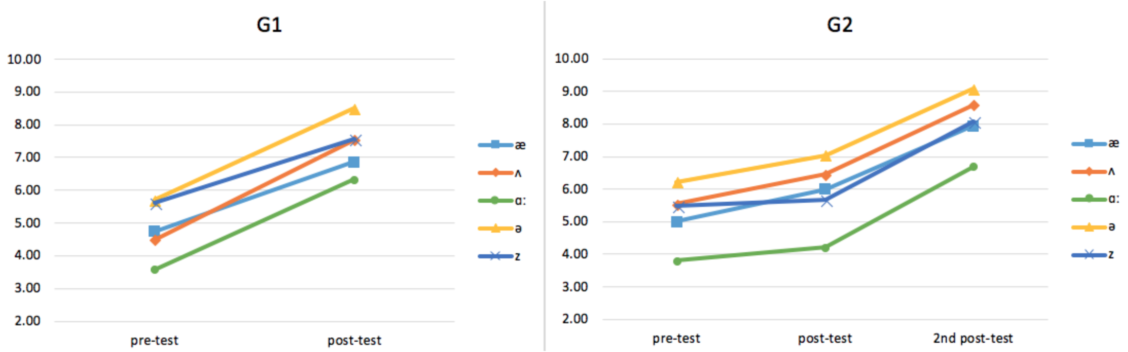


Figure 6. Mean scores for familiar items at pre-, post- and second post-test in the identification task

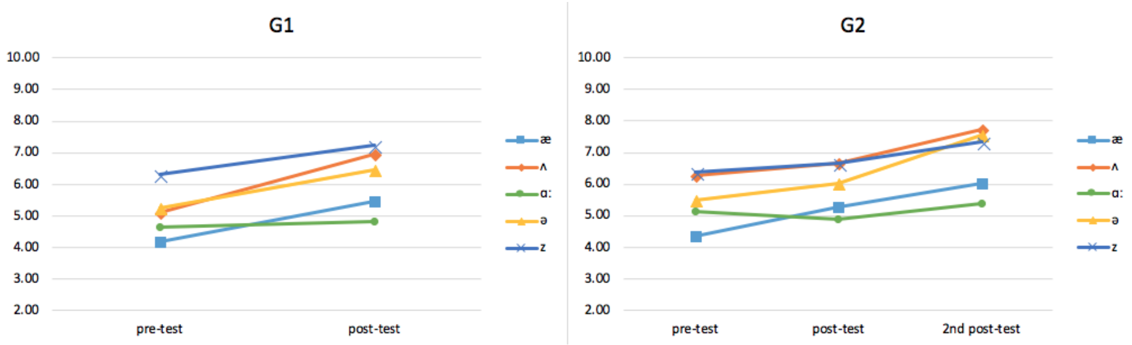


Figure 7. Mean scores for novel items at pre-, post- and second post-test in the identification task.

Production

As explained above, although G2 also acted as an experimental group for the perception data, it only acted as a control group for production to avoid imposing excessive demands on participants. Thus, the analyses of production data always compare the scores of G1 and G2C.

The ANOVA comparing both groups' scores across production tasks from pre- to post-test revealed a significant effect of the time variable (F(1,51) = 75.39, p = <0.001, r = 0.77) and a significant interaction between time and group (F(1,51) = 12.95, p = <0.001, r = 0.45) (Figure 8). A comparison of the scores obtained for individual sounds across the different production tasks revealed significant Time x Group interactions for /æ/ (F(1,51) = 4.41, p = 0.04, r = 0.28), /ʌ/ (F(1,51) = 7.12, p = 0.01, r = 0.35), /ɑː/ (F(1,51) = 5.66, p = 0.02, r = 0.32) and /ə/ (F(1,51) = 7.02, p = 0.01, r = 0.35), but not for /z/ (F(1,51) = 0.95, p = 0.33, r = 0.14). The mean scores, standard deviations, and 95% confidence intervals for each sound at pre- and post-tests for the different production tasks are presented in Appendix F. The analysis of the learners' performance in the different production tasks revealed significant differences between groups in the three tasks, although not for every sound (see Appendix D).

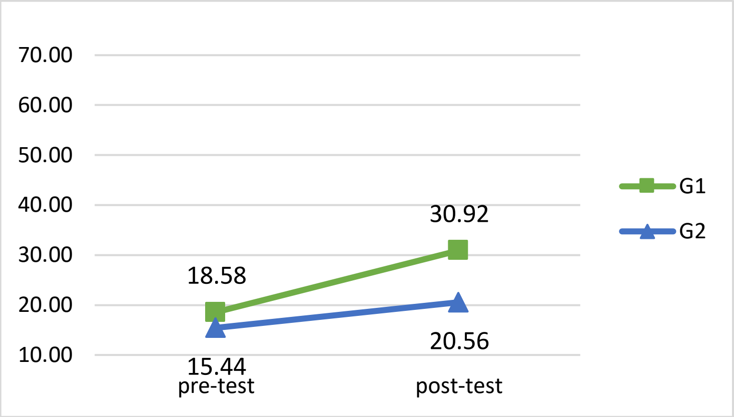


Figure 8. Mean scores across production tasks at pre- and post-test.

Regarding the imitation task, the analysis of the total scores from pre- to post-test did not reveal significant interactions between time and group (*F*(1,51) = 2.2, *p* = 0.14, *r* = 0.2). However, when considering the scores for each sound individually, a significant Time x Group interaction was found for /æ/ (mean scores illustrated visually in Figure 9).

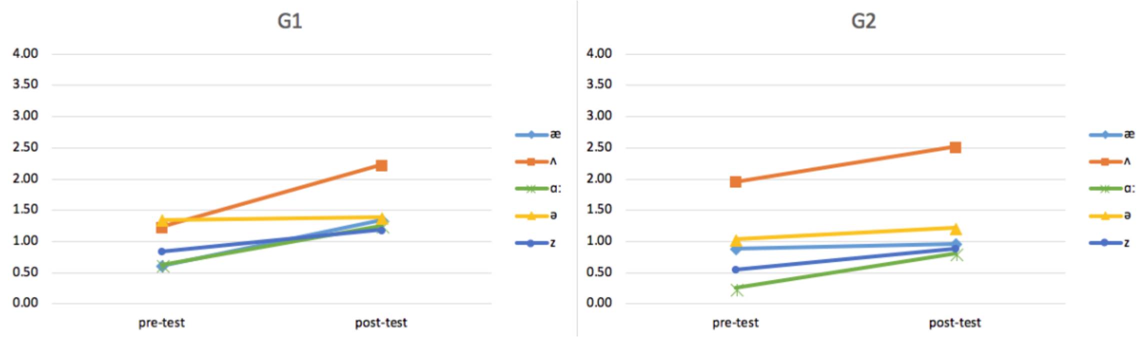


Figure 9. Mean scores in the imitation task at pre- and post-test

The analysis of the data from the sentence-reading task revealed a significant effect of time (*F*(1,51) = 47.69, *p* = <0.001, *r* = 0.7) and a significant interaction between time and group (*F*(1,51) = 9.8, *p* = 0.003, *r* = 0.4), which indicates that training exerted a positive impact in the learners' ability to pronounce the target sounds in controlled production. The improvements made by G1 were generally superior to those by G2 (Figure 10 and Figure 11). The analysis of the scores obtained for each sound in familiar items revealed significant interaction effects for /ɑː/ and /z/, but not for /æ/, /ʌ/ or /ə/ (Appendix D). The results for novel items reveal that the differences between groups were only significant for /ɑː/ and /ə/, but no significant interactions were found for /æ/, /ʌ/ or /z/.

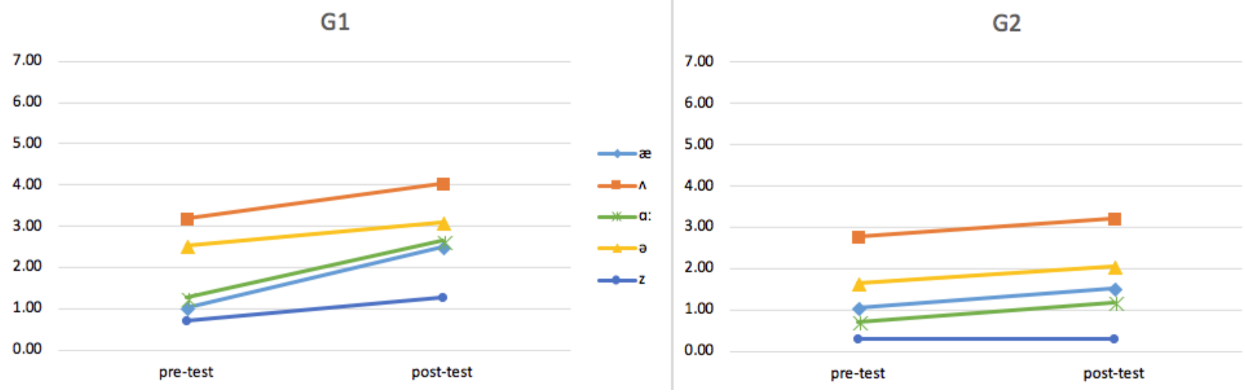


Figure 10. Mean scores for familiar stimuli in the sentence-reading task at pre- and post-test.

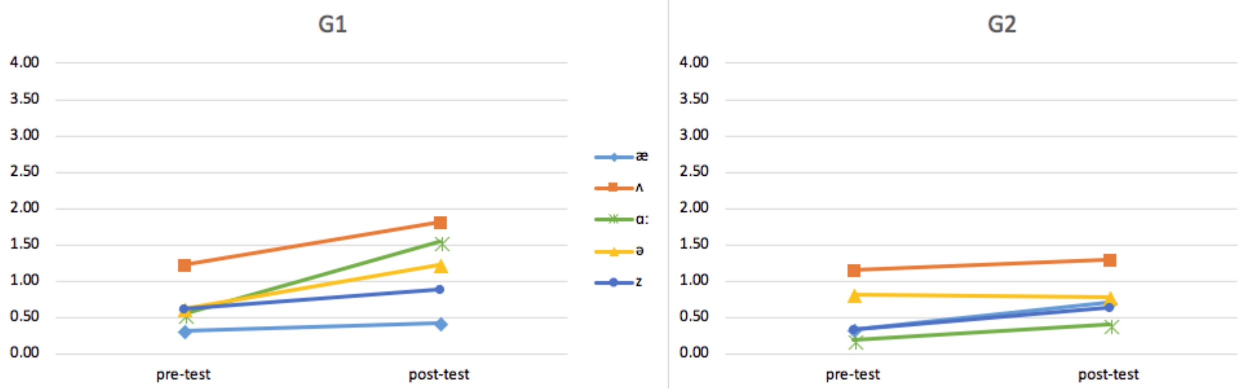


Figure 11. Mean scores for novel stimuli in the sentence-reading task at pre- and post-test.

Finally, the results from the total scores in the timed picture-description task revealed a significant effect of time (*F*(1,51) = 29.05, *p* = <0.001, *r* = 0.6) and a significant interaction between time and group (*F*(1,51) = 9.3, *p* = 0.004, *r* = 0.39). The analysis of the scores for individual sounds revealed significant Time x Group interactions for /ʌ/ and /ə/, but not for /æ/, /ɑː/ or /z/ (Appendix D). Both groups' mean scores on pre- and post-tests are illustrated visually in Figure 12.

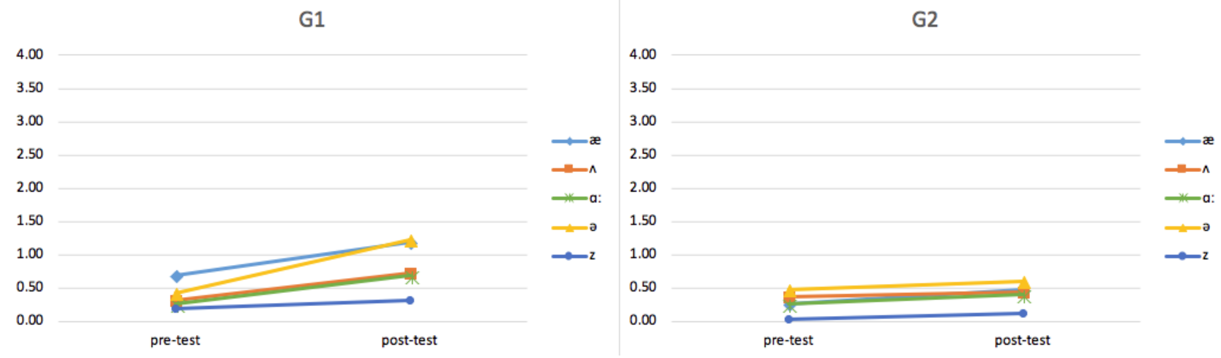


Figure 12. Mean scores in the timed picture-description task at pre- and post-test.

Discussion and Conclusions

The goal of this study was to explore the potential of the EFP app to help EFL learners improve their perception and production of a range of segmental features that tend to be fossilised in their interlanguage.

Given that these aspects tend to be difficult to modify without instruction, changes in the learners' perception and production should offer a clear measure of the learning potential of the app. In line with the data reported by Monroy-Casas (2001), the target aspects addressed also showed traits of fossilisation in the present study, as evidenced in the percentage of items participants mispronounced in the pre-test in the different production tasks (Table 2).

Table 2. *Percentage of Items that Were Mispronounced in the Pre-Test*

	Imitative		**Controlled**		**Spontaneous**	
	G1	**G2**	**G1**	**G2**	**G1**	**G2**
/æ/	84.6	77.8	88.5	88.1	82.7	93.5
/ʌ/	69.2	50.9	68.1	72.2	89.7	87.7
/a:/	84.4	93.5	87.3	93	93.3	93.5
/ə/	66.3	74.1	74.6	83.3	89.4	88
/z/	78.8	86.1	92.7	97	93.6	98.8

Note. For the sentence-reading task, only the familiar items have been considered.

The first research question addressed the potential of the EFP app to help learners improve their perception of the target features. While the differences between experimental and control groups did not reach statistical significance for every sound or every task, in general, the data show that training had a positive impact on the learners' perception of the target sounds, both in familiar and novel words. The results from the discrimination task show that both groups made similar improvements from pre- to post-test. However, the data from the identification task revealed significant differences between groups, both between G1 and G2C, as well as for G2E. This indicates that training was more effective in helping learners identify the target sounds correctly when they heard them individually than in helping them perceive differences among similar sounds when they were asked to compare three physically different tokens in triads of minimally paired words.

Focusing on the improvements made from pre- to post-test (G1 versus G2C) for familiar items in the identification task, training fostered significant differences between groups for four of the target sounds, namely /ʌ ɑː ə z/. Regarding the improvement made by G2E, the differences between the post-test and the second post-test reached the significance level for all the target sounds (/æ ʌ ɑː ə z/). As explained above, the voices used for testing and training were different. Thus, whenever the differences between groups reach the significance level, this can also be interpreted as generalisation to novel voices. Nevertheless, it is important to point out that, since the sample words the app offers for each sound are restricted to a relatively narrow set, it is difficult to ensure that improvements in familiar items are truly improvements in the learners' *perception* rather than them simply becoming familiar with the sounds with which these items are pronounced. Hence, the improvements made in novel items offer a more reliable measure of the learners' improvements in perception. Considering novel items, significant differences were found between G1 and G2C for /ʌ/, as well as between the post-test and second post-test scores for /ʌ ə z/ by G2E.

The second research question explored whether the perceptual training offered by the app could foster improvements in the participants' production of the target features. As in the perception tasks, the differences between groups did not reach the significance level for every sound or in every task. Nonetheless, the app-based training had a positive impact on the learners' pronunciation of the target features, both in familiar and novel words.

Considering the overall scores for individual sounds, the results show that training had a beneficial effect in the participants' production of /æ ʌ ɑː ə/. Nevertheless, the impact of instruction on the target sounds was not the same in every task. The results from the imitation task show that the differences between groups only reached significance for /æ/; the improvements made for the other sounds were modest and they were similar for both groups. In the sentence-reading task, a significant difference was found between groups when considering the total scores for the task, which indicates that training had a positive effect in the participants' controlled production of the target features. However, the analysis of the scores for individual sounds revealed that the differences between groups did not reach the significance level for every sound. In familiar items, the differences between groups were statistically significant for /ɑː/ and /z/; in novel items, the differences reached significance for /ɑː/ and /ə/. Finally, the results from the timed picture-description task revealed significant differences between groups when considering the total pre- and post-test scores, which indicates that training had a positive effect on the participants' spontaneous production of the target features. Nonetheless, the analysis of the scores for each sound individually revealed that the differences between groups were statistically significant only for /ʌ/ and /ə/.

In line with the results of other perceptual training studies (e.g. Bradlow et al., 1997; Carlet, 2017; Rato, 2013; Thomson, 2011), the data reported here shows that the perceptual training offered fostered improvements not only in the learners' perception of the target aspects addressed, but also in their production—even though the latter was not trained. The fact that this type of training helped learners improve their perception and production of these particular features is considered to be very positive, as they tend to be fossilised in the interlanguage of advanced learners of English and are therefore considered to be difficult to modify. Moreover, the improvements yielded by perceptual training were transferred to untrained words and to the participants' spontaneous production of some target sounds. The results suggest that using mobile applications for perceptual training can be particularly suitable to helping FL learners improve their perception and production of challenging pronunciation features. Training paradigms like this make it possible for learners work individually on aspects they find challenging. This can help learners improve their perception of the target features and foster the creation of adequate perceptual targets that guide their subsequent productions, which should, in turn, help them monitor their performance when engaged in communicative situations and facilitate autonomous work.

Following Plonsky and Oswald's (2014) recommended benchmarks of effect size for L2 research: small (r = 0.25, d = 0.60), medium (r = 0.4, d = 1.00) and large (r = 0.6, d = 1.40), the effect sizes in the data reported above are generally medium or large. The largest effect sizes are found in G2E's scores in the identification task, followed by G1's scores in the same task. The fact that training fostered more substantial improvements in the learners' perception than in their production of the target features is not surprising, as the type of training offered focused on perception and, in particular, identification. In general, the learners' scores after receiving instruction were far from the maximum scores in many of the tasks, especially in terms of production, which indicates that there is still much room for improvement. Nevertheless, it is important to consider that training added up to a total of less than 4 hours (approximately 20 mins a day over 10 days) and that stimuli variability was rather limited. As compared to HVPT paradigms which expose learners to a wide range of highly variable stimuli illustrating the pronunciation of the target sounds in different phonetic contexts by different voices, the training offered by the EFP app does not use a range of speakers and the target sounds are exemplified with a rather small number of sample words. Thus, although training did not foster significant differences between groups for all the target sounds or in every task, the results are encouraging.

It is important to highlight the fact that participants in both groups were receiving explicit instruction in phonetics at the time of the study. Given the role explicit instruction plays in FL learners' pronunciation (Saito, 2013), the fact that some participants in the control group showed some improvements from pre- to post-test is not surprising. Nonetheless, the fact that both groups were receiving the same classroom instruction is considered to offer a reliable measure of the effect exerted by the app, as the only difference between groups between testing times was the app-based training.

Additionally, research suggests that pronunciation training is more effective when form focused instruction is combined with explicit instruction (Saito, 2013). Hence, the explicit instruction participants were receiving in class may have also boosted the improvements fostered by the app, as learners had a lot of information about the target sounds and were familiarised with phonetic symbols, for example. However, research shows that perceptual training can facilitate learners' perception and production of challenging segments even when learners have not received much explicit pronunciation instruction (Thomson, 2011). Thus, although combinations of this type of perceptual training with explicit instruction may be especially beneficial, perceptual training paradigms like the one used in the EFP app should be suitable for any language learner, even if they do not have an explicit background in phonetics.

The fact that this app uses phonetic symbols to label the target sounds should help FL learners work on their perception and conceptualise the FL phonology more easily, as providing learners with a way of classifying sounds that does not rely on orthography should help them categorise sounds when they hear them and, consequently, facilitate the creation of adequate mental representations or "concepts" for those sounds (Mompean & Lintunen, 2015). One of the key elements to facilitate concept formation is using an adequate metalanguage which allows teachers and students to think and communicate adequately and precisely about the target concepts (Couper, 2011). Nevertheless, as Fraser (2006) notes, the metalanguage that we normally use to refer to speech is often strongly influenced by alphabetic writing which, in the case of English, makes pronunciation extremely complex given the lack of a one-to-one correspondence between phonemes and graphemes. In this regard, phonetic symbols are an orthography-independent way of representing speech which could act as some kind of metalanguage for learners. By simply offering learners labels with which to categorise the FL pronunciation, it should be easier for them to become aware that sounds they may have considered to be "similar" (e.g. /æ – ʌ/) are in fact different. As a case in point, if the only label Spanish EFL learners have for sounds occurring in the vowel space for the Spanish /a/ is the letter <a>, they will associate all the sounds that are articulated in that portion or near that portion of the vocal tract (/æ/, /ʌ/, /ɑː/ and, depending on the context, /ə/) to the mental representation they have for their native vowel /a/. Nonetheless, if learners realise that what they understood as /a/ can actually be four different English sounds (/æ ʌ/ and, depending on the context and the orthography, /ɑː ə/), even if they do not perceive differences among them at first, they should be better equipped to notice instances of these sounds in the input to which they are exposed and gradually become capable of categorising and producing them adequately.

It should be noted that the target features addressed in this study were considered relevant for this particular group of learners given their profile and because they were preparing for the C1 level. Because the goal of this study was to test the potential of the EFP app to help learners improve their perception and production of aspects that are difficult to modify, no ratings of accentedness or comprehensibility were done. However, the target features addressed are considered to be relevant from a functional load perspective (Brown, 1988; Catford, 1987). Empirical research investigating the theoretical predictions made by the functional load principle on a range of consonant sounds suggests that high functional load errors are more likely to impact the learners' comprehensibility and accentedness than low functional load errors (Munro & Derwing, 2006). To the researcher's knowledge, there are no studies exploring the impact of the above errors on Spanish EFL learners' intelligibility, comprehensibility, or accentedness. Nevertheless, even if EFL learners are considered to be generally intelligible at a B2 level, the functional load of the above features and the preliminary findings by Munro and Derwing (2006) with consonants suggest that recurrent mispronunciations of the above features are expected to have an impact on the learners' comprehensibility and accentedness. Additionally, failure to pronounce /ə/ adequately may completely alter the stress pattern of the word if pronounced as a full Spanish vowel, therefore altering one of the main cues for correct word identification (Field, 2005; Zielinski, 2008). Future research should continue to work along the lines of Munro and Derwing's (2006) study and explore the impact of this type of errors on the listeners' intelligibility, comprehensibility, and accentedness.

Given the challenging nature of pronunciation and the vast amount of practice FL learners need, the possibility to practise anywhere at any time is undoubtedly appealing. Provided that the audio quality is good and that the environment does not prevent learners from hearing the stimuli correctly, today's smartphones and tablets seem particularly suitable for different types of perceptual training, such as the one offered in the present study, approaches like HVPT, or for controlled production. Learners can practise on their own, at a time and place of their convenience, by using the devices they already have and use daily, which facilitates the integration of this type of training into the learners' routine, without having to sit at the computer or going to a computer lab.

Notwithstanding the above, and while app prices are often affordable, users may not always be willing to pay for apps. The EFP app cost 5,49€ at the time of the study.[7] While 26.5% (n=13) of participants who completed the post-test questionnaires in this study said that they would be willing to pay this price for the app, the majority said that they would not, 28.6% (n=14) because they considered it to be too expensive and 42.9% (n=21) because they never spend money on apps (Fouz-González, 2020).

The app explored in this study presents some limitations that could easily be overcome in future updates. Although the approach adopted can be useful to help learners conceptualise the sound system of English, if the app featured a wider range of sample words for each sound and more voices, training would undoubtedly be much more beneficial. In the trial run the researcher made in order to explore the app's stimuli, there were 12 words illustrating /æ/, 10 for /ɑː/ and /ʌ/, nine for /s/ and /z/, and 22 for /ə/. If users intend to use the app regularly and for long periods of time, training can be monotonous. Furthermore, the app does not offer users the possibility of choosing the sounds that they want to practise. Instead, users are presented with a randomised set of words featuring a range of English sounds. This can be useful to help users become familiarised with the sound system of English. Nevertheless, while studies using HVPT have shown that training with a wide set of sounds can be more beneficial than training with a small set (Nishi & Kewley-Port, 2007), given the limited variability of the stimuli in this app, presenting users with a randomised set of all the sounds featured in the app limits the amount of exposure to the target sounds considerably, as the sounds users may be interested in will only appear among many other words exemplifying other sounds.

CALL researchers have noted the need to address how different tools can be integrated in the curriculum for long periods of time (Burston, 2015; Chwo, Marek, & Wu, 2018). Nonetheless, for competences like pronunciation, which are not easily amenable to change without instruction after a certain age or experience level with the L1 (Pennington, 1998), short interventions can offer valuable insight into the potential of a given tool to foster changes in different aspects of FL learners' pronunciation. CAPT research has shown that there is no such thing as the ultimate tool capable of helping learners work on all their pronunciation needs (segmental and suprasegmental) and offer them accurate feedback on spontaneous production. Thus, technology should be understood as a facilitator that can enhance the way problematic aspects are presented and practised, allowing teachers and learners to use different tools for different purposes depending on their needs and the target features that they want to improve (Fouz-González, 2015). In this line, in the same way that CAPT research should not try to seek the perfect tool but to explore how different technologies can facilitate various types of training, when exploring apps, researchers should not try to find an app that can teach learners everything they need, but to explore the ways in which particular apps can support different aspects of pronunciation (Kaiser, 2018). In this respect, while the length of this study was relatively short because of the limited availability of participants and the desire to offer training to both groups, the results offer valuable information on the potential of this app to foster improvements in EFL learners' perception and production of features that are considered to be difficult to modify without instruction.

Acknowledgements

I would like to thank the journals' three anonymous reviewers for their extremely valuable comments on an earlier version of this paper. I am also grateful to Aurora González Vidal and Luis Prieto Valiente for their assistance with the statistical analyses, and to Pilar Mompean Guillamón, Ana Rosa Sánchez Muñoz and José Antonio Mompeán for their help with the evaluation of the participants' production. Finally, I am particularly grateful to the students who took part in this study and made it possible.

Notes

[1]. This has been calculated considering the responses of the 48 students who completed the initial

questionnaire.

2. Since the training learners received was perceptual, measuring perception was prioritised in this second post-test over measuring potential transfer to production.

3. Unlike the usual pattern in games, progression through the different "levels" here does not imply increasing difficulty. This is just a classification made by the researcher in order to quantify the number of stimuli learners were exposed to during training.

4. It should be noted that although the words *glasses* and *present* are considered as "familiar" stimuli because they were featured in the app, they only appeared in activity 2 (i.e. no audio input was offered – at least in the first 1000 stimuli of the trial run explained above).

5. Four of the items used to measure the learner's production were omitted from the analysis due to a mistake in the arrangement of stimuli in the production tests. They were *have* and *run* from the sentence-reading task, and *son* and *buys* from the timed picture-description task.

6. Even though the total number of participants in this study was 54, the pre-test data from three participants (13, 60 and 84) in the perception tests was lost due to a problem with the computers at university. Thus, neither pre- nor post-tests are considered in the analysis. Additionally, due to very poor audio quality in one of the recordings, the production data from one of the participants (47) was also omitted from the analysis. In all the analyses in this study, missing values have been excluded case-wise.

7. In this study, participants were given the app for free in exchange for their participation.

References

Best, C., & Tyler, M. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O. Bohn & M. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). Amsterdam, The Netherlands: John Benjamins.

Bradlow, A. R., Pisoni, D. B., Akahana-Yamada, R., & Tohkura, Y. (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*, 2299–2310.

Brown, A. (1988). Functional load and the teaching of pronunciation. *TESOL Quarterly*, *22*(4), 593–606.

Burston, J. (2015). Twenty years of MALL project-implementation: A meta-analysis of learning outcomes. *ReCALL*, *27*(1), 4–20.

Carlet, A. (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 doctoral dissertation). Universitat Autònoma de Barcelona, Barcelona, Spain.

Catford, J.C. (1987). Phonetics and the teaching of pronunciation. In J. Morley (Ed.), *Current perspectives on pronunciation* (pp. 87–100). Washington, DC: TESOL.

Chwo, S. M. G., Marek, M. W., & Wu, W-C. V. (2018). Meta-analysis of MALL research and design. *System*, *74*, 62–72

Colpaert, J. (2004). From courseware to coursewear? *Computer Assisted Language Learning*, *17*(3–4), 261–266.

Council of Europe. (2018). *Common European framework of reference for languages: Learning, teaching, assessment. Companion volume with new descriptors.* Retrieved from: https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Couper, G. (2011). What makes pronunciation teaching work? Testing for the effect of two variables: socially constructed metalanguage and critical listening. *Language Awareness*, *20*(3), 159–182.

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, *39*, 399–423.

Flege, J. E. (1995). Second-language speech learning: Theory, findings and problems. In W. Strange (Ed.), *Speech perception and linguistic experience* (pp. 233–277). Timonium, MD: York Press Inc.

Foote, J., & McDonough, K. (2017). Using shadowing with mobile technology to improve L2 pronunciation. *Journal of Second Language Pronunciation, 3*(1), 34–56.

Fouz-González, J. (2015). Trends and directions in computer assisted pronunciation training. In J. Mompean & J. Fouz-González (Eds.), *Investigating English pronunciation: Trends and directions* (pp. 314–342). Basingstoke, UK/New York, NY: Palgrave Macmillan.

Fouz-González, J. (2019). Podcast-based pronunciation training: Enhancing FL learners' perception and production of fossilised segmental features. *ReCALL, 31*(2), 150–169. https://doi.org/10.1017/S0958344018000174

Fouz-González, J. (2020). Using the English File Pronunciation app for pronunciation training: The learners' views. In A. Bocanegra-Valle (Ed.), *Applied linguistics and knowledge transfer: Employability, internationalization, and social challenges* (pp. 77–103). Bern: Peter Lang.

Fraser, H. (2006). Helping teachers help students with pronunciation: A cognitive approach. *Prospect, 21*(1), 80–96.

Gómez-Lacabex, E., García-Lecumberri, M. L., & Cooke, M. (2008). Identification of the contrast full vowel-schwa: Training effects and generalization to a new perceptual context. *Ilha do Desterro, 55*, 173–196.

Hardison, D.M. (2004). Generalization of computer-assisted prosody training: Quantitative and qualitative findings. *Language Learning & Technology, 8*(1), 34–52. Retrieved from: https://scholarspace.manoa.hawaii.edu/bitstream/10125/25228/1/08_01_hardison.pdf

Hualde, J. I. (2014). *Los sonidos del español*. Cambridge, UK: Cambridge University Press.

Kaiser, D. (2018). Mobile-assisted pronunciation training: The iPhone pronunciation app project. *IATEFL Pronunciation Special Interest Group Journal*, *58*, 38–52.

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics, 36*(3), 345–366.

Levis, J. (2007). Computer technology in teaching and researching pronunciation. *Annual Review of Applied Linguistics, 27,* 184–202.

Liakin, D., Cardoso, W., & Liakina, N. (2015). Learning L2 pronunciation with a mobile speech recognizer: French /y/. *CALICO Journal*, *32*(1), 1–25.

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

Lord, G. (2008). Podcasting communities and second language pronunciation. *Foreign Language Annals, 41*, 374–389.

Mompean, J.A., & Lintunen, P. (2015). Phonetic notation in foreign language teaching and learning: Potential advantages and learners' views. *Research in Language*, *13*(3), 292–314.

Monroy-Casas, R. (2001). Profiling the phonological processes shaping the fossilised IL of adult Spanish learners of English. Some theoretical implications. *International Journal of English Studies, 1,* 157–217.

Motohashi-Saigo, M., & Hardison, D.M. (2009). Acquisition of L2 Japanese geminates training with waveform displays. *Language Learning & Technology*, *13*(2), 29–47. Retrieved from: https://scholarspace.manoa.hawaii.edu/bitstream/10125/44179/1/13_02_motohashisaigohardison.pdf

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, *34*, 520–531.

Neri, A., Cucchiarini, C., & Strik, H. (2008) The effectiveness of computer-based corrective feedback for improving segmental quality in L2-Dutch. *ReCALL, 20*(2), 225–243.

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

O'Brien, M. G. (2011). Teaching and assessing pronunciation with computer technology. In N. Arnold & L. Ducate (Eds.), *Present and future promises of CALL: From theory and research to new directions in language teaching* (pp. 375–406). San Marcos, TX: CALICO.

Olson, D. J. (2014). Benefits of visual feedback on segmental production in the L2 classroom. *Language Learning & Technology*, *18*(3), 173–192. Retrieved from: https://scholarspace.manoa.hawaii.edu/bitstream/10125/44389/1/18_03_olson.pdf

Oxford University Press. (2012). *English File Pronunciation*. (Version 1.1). [Mobile application software]. Retrieved from: https://itunes.apple.com/es/app/english-file-pronunciation/id520767531?mt=8

Pennington, M.C. (1998). The teachability of pronunciation in adulthood: A reconsideration. *International Review of Applied Linguistics, 36*, 323–41.

Pennington, M.C. (1999). Computer-aided pronunciation pedagogy: Promise, limitations, directions. *Computer-Assisted Language Learning, 12*(5), 427–440.

Plonsky, L., & Oswald, F. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912.

Qian, M., Chukharev-Hudilainen, E., & Levis, J. (2018). A system for adaptive high-variability segmental perceptual training: Implementation, effectiveness, transfer. *Language Learning & Technology*, *22*(1), 69–96. Retrieved from: https://scholarspace.manoa.hawaii.edu/bitstream/10125/44582/1/22_01_qianchukharev-hudilainenlevis.pdf

Ramírez-Verdugo, D. (2006). A study of intonation awareness and learning in non-native speakers of English. *Language Awareness, 15*(3), 141–159.

Rato, A. (2013). *Cross-language perception and production of English vowels by Portuguese learners: The effects of perceptual training* (Unpublished doctoral dissertation). Universidade do Minho, Braga, Portugal.

Rato, A., Rauber, A. S., Kluge, D. C., & Santos, G. R. (2015). Designing speech perception tasks with TP. In J. A. Mompean & J. Fouz-González (Eds.), *Investigating English pronunciation: Trends and directions* (pp. 295–313). Basingstoke, UK: Palgrave Macmillan.

Saito, K. (2013). Re-examining effects of form-focused instruction on L2 pronunciation development: The role of explicit phonetic information. *Studies in Second Language Acquisition, 35*, 1–29.

Thomson, R. (2011). Computer assisted pronunciation training: Targeting second language vowel perception improves pronunciation. *CALICO Journal*, *28,* 744–765.

Thomson, R. (2018). *English Accent Coach* [Computer program]. Version 2.3. Retrieved from www.englishaccentcoach.com

Thomson, R., & Derwing, T. D. (2015). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics, 36*(3), 326–344.

Uther, M., Uther, J., Athanasopoulos, P., Singh, P., & Akahane-Yamada, R. (2007). Mobile adaptive CALL (MAC). A lightweight speech-based intervention for mobile language learners. *Proceedings of INTERSPEECH 2007*, 2329–2332.

Zielinski, B. W. (2008). The listener: No longer the silent partner in reduced intelligibility. *System*, *36*, 69–84.

Appendix A. Testing stimuli for the discrimination task

/æ – ʌ/		/æ – ɑː/		/ʌ – ɑː/		/ə/	/s – z/	
hat	hut	cap	carp	cup	carp	and	sap	zap
bag	bug	hat	heart	hut	heart	but	seal	zeal
bat	but	ban	barn	bun	barn	that	sing	zing
matt	mutt	pack	park	come	calm	at	said	z
cap	cup	pat	part	buck	bark	have	racing	raising
track	truck	cad	card	huff	half	must	fussy	fuzzy
app	up	hack	hark	pus	pass	than	muscle	muzzle
ban	bun	am	arm	done	darn	some	precedent	president
bank	bunk	bat	Bart	cluck	clerk	can	bus	buzz
pat	putt	chat	chart	dunce	dance	does	price	prize

Distractors /s - ʃ/					
sue	shoe	seesaw	seashore	Iris	Irish
see	she			mass	mash

Appendix B. Testing stimuli for the identification task

Familiar						Distractors
/æ/	/ʌ/	/ɑː/	/ə/	/s/	/z/	/ʃ/
stamp	cousin	argue	famous	books	flies	shoe
capital	uncle	aunt	Africa	works	buys	she
happen	young	car	ago	nice	reads	issue
actor	comfortable	dark	picture	writes	museum	mission
flat	hundred	answer	dangerous	cooks	music	bush
garage	son	dance	dinner	costs	exams	
back	stomach	garden	October	eats	please	
bad	under	star	pilot	speaks	arrives	
black	bus	class	second	lettuce	words	
have	come	glasses	sugar	police	present	

Novel					
/æ/	**/ʌ/**	**/ɑː/**	**/ə/**	**/s/**	**/z/**
brag	jump	farm	manner	ceiling	zombie
jazz	money	park	hospital	centre	zebra
flag	blood	past	another	December	busy
gang	tough	smart	against	decent	easy
happy	couple	laugh	problem	discipline	fizzy
fax	run	demand	forget	recipe	amazing
hang	lunch	dart	prison	peace	rose
glad	touch	bark	horizon	niece	amuse
gas	multiple	balm	brother	ice	these
fan	sun	garlic	family	pace	cheese

Appendix C. Stimuli for the imitation task, sentence-reading task and timed picture-description task

Imitation task

/æ/	/ɑː/	/ʌ/	/ə/	/z/
hat	heart	cup	manner	zeal
bag	barn	buck	hospital	z
cap	park	huff	problem	fuzzy
track	chart	cluck	prison	prize

Sentence-reading task

Familiar				
/æ/	**/ʌ/**	**/ɑː/**	**/ə/**	**/z/**
stamp	cousin	argue	famous	flies
capital	uncle	aunt	Africa	buys
happen	young	car	ago	reads
actor	comfortable	dark	picture	museum
flat	hundred	answer	dangerous	music
garage	son	dance	dinner	exams
back	stomach	garden	October	please
bad	under	star	pilot	arrives
black	bus	class	second	words
have	come	glasses	sugar	present

Novel				
/æ/	**/ʌ/**	**/ɑː/**	**/ə/**	**/z/**
fan	gum	far	father	frozen
hang	bug	father	lemon	magazine
lack	run	half	oven	zebra
sad	brother	large	student	lazy
anger	fun	guitar	camera	size

Timed picture-description task

/æ/	/ʌ/	/ɑː/	/ə/	/z/
stamp	cousin	aunt	famous	music
capital	uncle	car	Africa	buys
actor	young	dark	picture	flies
black	son	dance	dangerous	museum

Appendix D. Results from the statistical analyses for individual sounds in the identification task and in the production tasks

Identification Tasks

	Familiar stimuli	
Target	**G1**	**G2E**
/æ/	$F(1,44) = 2.72$, $p = 0.11$, $r = 0.24$	$Z = 3.72$, $p = <0.001$, $r = 0.83$*
/ʌ/	$F(1,44) = 10.31$, $p = 0.002$, $r = 0.44$*	$Z = 3.55$, $p = <0.001$, $r = 0.79$*
/ɑː/	$F(1,44) = 18.98$, $p = <0.001$, $r = 0.55$*	$t(19) = 6.24$, $p = <0.001$, $d = 1.39$*
/ə/	$F(1,44) = 7.47$, $p = 0.009$, $r = 0.38$*	$Z = 3.08$, $p = 0.002$, $r = 0.69$*
/z/	$F(1,44) = 8.84$, $p = 0.005$, $r = 0.41$*	$Z = 3.43$, $p = 0.001$, $r = 0.77$*

	Novel stimuli	
Target	**G1**	**G2E**
/æ/	$F(1,44) = 0.38$, $p = 0.54$, $r = 0.09$	$t(19) = 1.32$, $p = 0.2$, $d = 0.29$
/ʌ/	$F(1,44) = 4.73$, $p = 0.03$, $r = 0.31$*	$Z = 2.7$, $p = 0.007$, $r = 0.6$*
/ɑː/	$F(1,44) = 0.45$, $p = 0.5$, $r = 0.1$	$Z = 0.865$, $p = 0.387$, $r = 0.19$
/ə/	$F(1,44) = 1.04$, $p = 0.31$, $r = 0.15$	$t(19) = 3.05$, $p = 0.007$, $d = 0.68$*
/z/	$F(1,44) = 1.63$, $p = 0.21$, $r = 0.19$	$Z = 2.76$, $p = 0.006$, $r = 0.62$*

Production tasks

Target	Imitation task
/æ/	$F(1,51) = 5.95$, $p = 0.02$, $r = 0.32$*
/ʌ/	$F(1,51) = 1.65$, $p = 0.2$, $r = 0.18$

/ɑː/	*F*(1,49) = 0.07, *p* = 0.8, *r* = 0.04
/ə/	*F*(1,51) = 0.58, *p* = 0.44, *r* = 0.11
/z/	*F*(1,51) = 0.002, *p* = 0.96, *r* = 0.01

	Sentence-reading task	
Target	**Familiar**	**Novel**
/æ/	*F*(1,51) = 3.69, *p* = 0.06, *r* = 0.26	*F*(1,51) = 1.76, *p* = 0.19, *r* = 0.18
/ʌ/	*F*(1,51) = 1.36, *p* = 0.24, *r* = 0.16	*F*(1,51) = 3.68, *p* = 0.06, *r* = 0.26
/ɑː/	*F*(1,51) = 4.41, *p* = 0.04, *r* = 0.28*	*F*(1,51) = 6.32, *p* = 0.015, *r* = 0.33*
/ə/	*F*(1,51) = 0.3, *p* = 0.58, *r* = 0.08	*F*(1,51) = 11.45, *p* = 0.001, *r* = 0.43*
/z/	*F*(1,51) = 4.21, *p* = 0.04, *r* = 0.28*	*F*(1,51) = 0.01, *p* = 0.9, *r* = 0.02

Timed picture-description task	
Target	
/æ/	*F*(1,51) = 1.57, *p* = 0.21
/ʌ/	*F*(1,51) = 4.07, *p* = 0.04, *r* = 0.27*
/ɑː/	*F*(1,51) = 2.15, *p* = 0.14, *r* = 0.2
/ə/	*F*(1,51) = 8.93, *p* = 0.004, *r* = 0.39*
/z/	*F*(1,51) = 0.25, *p* = 0.61, *r* = 0.07

Note: The results from the ANOVAs comparing G1 and C2C (in the identification task and in all the production tasks) only report the interactions between the time and group variables. Statistically significant results are marked with an asterisk.

Appendix E. Mean scores, standard deviations and 95% Confidence Intervals (CI) for each sound in the identification task at pre-test (pre), post-test (post) and second post-test (post2)

Note: Due to the size of the table in Appendix E, the table is presented here.

Appendix F. Mean scores, standard deviations (in brackets) and 95% Confidence Intervals (CI) for each sound at pre- and post-tests for the different production tasks

Imitation task

		pre			**post**		
Sound	**Group**	***M***	***SD***	***CI***	***M***	***SD***	***CI***
/æ/	G1	0.6	0.9	0.2,1	1.3	1.5	0.8,1.8
	G2	0.9	1.2	0.5,1.5	1	1.1	0.5,1.5
/ʌ/	G1	1.2	1	0.8,1.7	2.2	1.3	1.6,2.8
	G2	2	1.2	1.5,2.4	2.5	1.6	1.9,3.1
/ɑː/	G1	0.6	1	0.3,0.9	1.3	1.3	0.8,1.7
	G2	0.3	0.5	0,0.6	0.8	0.9	0.4,1.2

/ə/	G1	1.3	1.2	0.9,1.8	1.4	1.2	0.9,1.9
	G2	1	1.1	0.6,1.5	1.2	1.4	0.7,1.7
/z/	G1	0.8	1.2	0.5,1.2	1.2	1.2	0.7,1.7
	G2	0.6	0.8	0.2,0.9	0.9	1.3	0.4,1.4

Note: The maximum score for each sound was 4.

Sentence-reading task

		Familiar						**Novel**					
		pre			**post**			**pre**			**post**		
Sound	**Group**	***M***	***SD***	***CI***	***M***	***SD***	***CI***	***M***	***SD***	***CI***	***M***	***SD***	***CI***
/æ/	G1	1.04	1.34	0.4,1.7	2.5	2.55	1.5,3.5	0.31	0.54	0.1,0.5	0.42	0.75	0,0.8
	G2	1.07	1.92	0.4,1.7	1.52	2.65	0.5,2.5	0.33	0.67	0.1,0.6	0.7	1.13	0.3,1.1
/ʌ/	G1	3.19	1.65	2.6,3.8	4.04	2.16	3.3,4.8	1.23	0.81	0.9,1.5	1.81	1.02	1.4,2.1
	G2	2.78	1.45	2.2,3.4	3.22	1.64	2.5,4	1.15	0.77	0.8,1.4	1.3	0.77	0.9,1.6
/ɑː/	G1	1.27	1.51	0.7,1.8	2.65	2.33	1.8,3.5	0.54	1.1	0.2,0.9	1.54	1.5	1.1,2
	G2	0.7	0.95	0.2,1.2	1.19	1.71	0.4,2	0.19	0.48	-0.1,0.5	0.41	0.79	-0.1,0.9
/ə/	G1	2.54	2.25	1.7,3.4	3.12	2.53	2.2,4	0.62	0.89	0.2,1	1.23	1.14	0.8,1.7
	G2	1.67	2.03	0.8,2.5	2.07	2.07	1.2,3	0.82	1.14	0.4,1.2	0.78	1.08	0.3,1.2
/z/	G1	0.73	2.03	0.1,1.3	1.27	2.55	0.5,2	0.62	1.41	0.2,1	0.89	1.6	0.3,1.5
	G2	0.3	0.61	-0.3,0.9	0.3	0.67	-0.4,1	0.33	0.62	-0.1,0.7	0.63	1.27	0.1,1.2

Note: The maximum scores for Familiar items were 10 for /ʌ ɑː ə z/ and 9 for /æ/. For Novel items, the maximum scores were 5 for /æ ɑː ə z/ and 4 for /ʌ/.

Timed picture-description task

		pre			post		
		M	***SD***	***CI***	***M***	***SD***	***CI***
/æ/	G1	0.69	0.92	0.4,1	1.19	1.35	0.8,1.6
	G2	0.26	0.59	0,0.6	0.48	0.8	0.1,0.9
/ʌ/	G1	0.31	0.54	0.1,0.5	0.73	0.82	0.4,1
	G2	0.37	0.62	0.1,0.6	0.44	0.69	0.1,0.7
/ɑː/	G1	0.27	0.72	0,0.5	0.69	0.88	0.4,1
	G2	0.26	0.44	0,0.5	0.41	0.74	0.1,0.7
/ə/	G1	0.42	0.85	0.1,0.8	1.23	1.3	0.8,1.6
	G2	0.48	0.89	0.1,0.8	0.59	0.84	0.2,1
/z/	G1	0.19	0.69	0,0.4	0.31	0.73	0.1,0.5
	G2	0.04	0.19	-0.2,0.2	0.11	0.42	-0.1,0.3

Note: The maximum scores were 4 for /æ ɑː ə/ and 3 for /ʌ z/.

About the Author

Jonás Fouz-González holds a PhD in English Applied Linguistics. He is a lecturer and researcher at the University of Murcia, where he teaches English Phonetics and several EFL courses. His research interests are English phonetics and phonology, second language acquisition, computer-assisted pronunciation training, and mobile-assisted language learning.

E-mail: jfouz@um.es