Studies in Second Language Acquisition (2025), **47**, 1044–1070
doi:10.1017/S0272263125101083

RESEARCH ARTICLE

Does timing matter?

Exploring different windows of maximal opportunity to enhance the effectiveness of high variability phonetic training

Charlie Nagle[1], Jose A Mompean[2] and Jonás Fouz-González[3]

[1]Department of Spanish and Portuguese, The University of Texas at Austin, Austin, TX, USA; [2]Department of English, University of Murcia, Murcia, Spain and [3]Department of Didactics of Language and Literature, University of Murcia, Murcia, Spain
Corresponding author: Charlie Nagle; Email: cnagle@austin.utexas.edu

(Received 30 September 2024; Revised 20 May 2025; Accepted 17 June 2025)

Abstract
A large body of literature has examined perceptual training, especially using the high variability phonetic training (HVPT) technique, where multiple talkers are included in the training set to help learners develop more accurate additional (second) language (L2) speech sound categories. Yet, most experimental studies focus on relatively short-term gains using a pre-post–delayed design, providing limited insight into longer-term training effects and how the timing of training might regulate its effectiveness. To begin addressing this gap, we implemented HVPT at two contextually relevant windows of opportunity during a university study program. Thirty-six first (native) language Spanish students participated in this study. Students were randomly assigned to two groups. One group (G1) received training at the beginning of their study program, which coincided with the onset of intensive L2 exposure; the second group (G2) received training in the second year, while enrolled in an English phonetics and phonology course. Both groups completed four HVPT sessions (identification tasks) focusing on a set of challenging L2 English vowels (/iː ɪ æ ʌ ɜː e ɒ ɔː/). Perception was measured at four testing times (in years 1 and 2, before and after HVPT) with identification tasks. The results showed that HVPT had a positive impact regardless of the timing of its implementation. However, students also improved outside of training, which suggests that intensive language study can facilitate some perceptual learning.

Keywords: HVPT; L2 pronunciation; longitudinal research; window of maximal opportunity

Introduction

Learning to perceive the sounds of an additional (second) language (L2) is a major challenge for adult learners because, by the time L2 learning begins, perception has become attuned to and optimized for the sounds and contrasts present in the first

(native) language (L1) (Kuhl et al., 2007). Perceptual attunement is advantageous for L1 perception but poses a challenge for L2 learning because the L1 may act as a filter through which L2 sounds are perceived. Several models have been proposed to account for such challenges (e.g., Best & Tyler, 2007; Flege & Bohn, 2021; van Leussen & Escudero, 2015). Despite their theoretical differences, all models converge on the fact that assimilation patterns—that is, the ways in which L2 sounds are mapped onto and potentially equated with L1 categories—shape the nature and difficulty of the learning task. Models are also in agreement about the fact that phonetic learning, including the creation of new perceptual categories, remains possible throughout the lifespan, but the extent to which learning occurs and the time course over which it occurs depend on myriad factors, including how well L2 sounds map onto L1 categories, how productive L2 contrasts are, and the quantity and quality of input that learners receive, among others.

A common challenge for most instructed learners is that their exposure to input is limited, at least compared to learners who are living and working in a context where the L2 is spoken. Thus, in an instructed context, targeted perceptual training has the potential to catalyze and/or accelerate perceptual learning, leading to marked gains in learners' ability to discriminate L2 contrasts and identify minimally contrastive L2 words. Research has unequivocally demonstrated that perceptual training is effective for perceptual learning and can even have a spillover effect on production (Sakai & Moorman, 2018; Uchihara et al., 2024). In their meta-analysis of perception training studies, Sakai and Moorman (2018) reported that training leads to a medium gain in perception accuracy. Another more recent meta-analysis focusing on single-talker versus multitalker perceptual training demonstrated that multitalker conditions provide a small but reliable advantage compared to their single-talker counterparts (Zhang, Cheng & Zhang, 2021). Up to this point, researchers working in this area have typically manipulated the nature of the training that learners receive, whereas other variables, such as when the training is administered, have not been systematically explored. Yet, issues of timing are critical to enhancing the applicability of research findings to the classroom because timing is a variable that instructors can alter.

The high variability phonetic training (HVPT) paradigm is an ideal testing ground for investigating the effects of timing because there is a large body of research pointing to the benefits of this type of multitalker training. As a result, the current research landscape is positioned to move beyond demonstrating that HVPT is effective, turning instead to questions about how it can be optimized. In this study, we investigated how the timing of HVPT affects its efficacy, comparing two groups who received training at distinct moments in their instructional trajectories. An added benefit of examining the timing of training is that timing studies demand a more comprehensive period of observation, encompassing the onset and offset of training for multiple groups. This type of study can therefore provide additional insight into the longer-term, longitudinal impact of training, when training is incorporated into the curriculum at different points.

Background

Perceptual training

Perception can be trained using a variety of paradigms, but HVPT has become the dominant technique in the literature. In HVPT, multiple talkers and phonetic contexts

https://doi.org/10.1017/S0272263125101083 Published online by Cambridge University Press

are trained simultaneously to encourage the development of robust and generalizable perceptual categories (i.e., categories that are not context- or speaker-dependent). Typically, HVPT is implemented as a forced-choice identification task with feedback, although discrimination tasks and other modifications of the canonical HVPT format —such as incorporating production practice—have also been used (for an overview of HVPT methods, see Cebrian, Gavaldà, Gorba & Carlet, 2024; Thomson, 2018). During identification training, listeners hear an auditory stimulus (typically a real word or nonsense word) and are asked to identify the sound or pattern that they hear from a set of options. If listeners respond correctly, they get a positive indication and move on to the next trial. If they answer incorrectly, they get a negative indication and are shown what the correct answer was. Sometimes, the trial is repeated so that they can select the correct answer.

The defining characteristic of HVPT is a multitalker approach, which has been shown to lead to small but reliable gains when compared to single-talker training paradigms (Zhang et al., 2021). The multitalker benefit dates back to the very origin of HVPT research, where across a series of studies Logan, Lively and Pisoni (1991) and Lively, Logan & Pisoni (1993) examined how well L1 Japanese speakers learned the English /l/-/ɹ/ contrast, considering both their performance from pretest to posttest and their ability to generalize to untrained talkers (i.e., to voices that were not included in the training set). Their results suggested a benefit for the multitalker approach, and, since these seminal studies, a large body of research has shown that HVPT is beneficial for training a range of segmental (e.g., Carlet & Cebrian, 2022; Fouz-González & Mompean, 2021; Thomson, 2012) and suprasegmental (e.g., Perrachione, Lee, Ha & Wong, 2011; Silpachai, 2020; Tremblay, Kim, Kim & Cho, 2023; Wang, Spence, Jongman & Sereno, 1999) features.

Set against this backdrop, researchers have begun to examine a range of training-related variables that could moderate how effective HVPT is, including the response options given during training (Fouz-González & Mompean, 2021), the training task used (Cebrian et al., 2024), and the type of stimuli used and the conditions in which they are presented (Mora, Ortega, Mora-Plaza & Aliaga-García, 2022). Intense interest in HVPT has also triggered a reexamination of the multitalker advantage, which may not be as robust as initially hypothesized. A recent replication of the Logan et al. (1991) and Lively et al. (1993) studies by Brekelmans, Lavan, Saito, Clayards & Wonnacott (2022), shoring up several methodological shortcomings present in the original design, found no clear evidence for a multitalker benefit. At the same time, the authors recognized the importance of continuing to investigate "how and under what circumstances variability can support the efficacy of phonetic training" (p. 21). Certainly, contemporary HVPT studies are beginning to shed light on this issue, but one critical but overlooked variable that might affect how much learners benefit from HVPT is its timing, or the point at which it is integrated into language instruction.

Time and timing in L2 speech learning

In their review of longitudinal research in L2 acquisition, Ortega and Iberri-Shea observed that "many questions concerning L2 learning are fundamentally questions of time and timing" (2005, p. 27). Applied speech researchers have made similar observations, noting that longitudinal research on time and timing can play a special role in shaping instructional targets. For example, reflecting upon goals for pronunciation instruction, Derwing posed the following question: "[W]ouldn't it be helpful to have

https://doi.org/10.1017/S0272263125101083 Published online by Cambridge University Press

some longitudinal studies to know which aspects of pronunciation will likely take care of themselves over time? Such information would allow teachers to focus on more intransigent problems" (2010, p. 27). A logical extension to this question is understanding how the timing of instruction affects learning, with the goal of optimizing both the *what* and the *when* of instruction (i.e., the target structures that should be prioritized in training and the moment at which the training should be implemented).

Two types of studies in particular offer valuable insights into L2 speech learning: longitudinal observational studies, which provide insight into how pronunciation develops in the absence of instruction; and longitudinal experimental studies, which shed light on the longitudinal impact of instruction. In longitudinal observational studies, learners do not receive any pronunciation-specific instruction. Instead, pronunciation is tracked over time to examine changes that occur during the natural course of learning, be it naturalistic, instructed, or both. Despite a growing body of longitudinal studies of this type (Nagle, 2021a), longitudinal studies of speech perception are surprisingly rare, but the existing literature shows that trajectories are often complex and non-linear (for L2 stops, e.g., Casillas, 2020; Nagle, 2018, 2021b; for L2 vowels, e.g., Kim, Clayards & Goad, 2018).

In terms of an experimental approach, there have been several HVPT studies demonstrating retention of learning one (Fouz-González & Mompean, 2021; Thomson, 2012), two (Carlet & Cebrian, 2022; Rato, 2014), and even six months (Lively, Pisoni, Yamada, Tohkura & Yamada, 1994; Silpachai, 2020) after training, but those studies are all conceptualized as examining retention on a delayed posttest without necessarily considering the impact of instruction within a more comprehensive developmental window. One notable exception is the SALA project, the goal of which was to examine instructed L2 English learners' language development longitudinally, as a result of general formal instruction (FI) with a focus on grammar, reading, and writing (i.e., without a focus on pronunciation) and study abroad (for an overview, see Pérez-Vidal, 2014). As part of the project, Mora (2014) reported that learners' perception of English consonant and vowel contrasts improved significantly after general FI but not after study abroad (SA). Reflecting upon the results of the study, Mora hypothesized that "the outcome of the SA period might have been very different in the domain of phonology had the participants received specific focused instruction on pronunciation skills or intensive perceptual training on the sound contrasts of English during FI" (2014, p. 190). In other words, if participants had received pronunciation instruction immediately before study abroad, they may have been better prepared to benefit from the extensive, variable phonetic input available during their immersion experience, in which case additional gains in perceptual ability may have been evident after study abroad (see also Lord, 2010). In summary, the timing of targeted pronunciation instruction relative to other instructional and developmental milestones may play a critical role in shaping its effectiveness.

Crucially, if timing is to be manipulated, then potential training windows must be carefully evaluated considering the learning context. Put another way, out of an infinite set of potential timing options, researchers need a useful heuristic for selecting and evaluating options that are likely to be impactful. Otherwise, it is unclear why one moment would be preferred over any other. Mora's idea of synergistic effects can provide such a guide. In this study, we integrate HVPT into a university degree program at the onset of intensive FI and in a pronunciation-related course. We consider these two options potential windows of maximal opportunity (WMOs; Derwing & Munro,

https://doi.org/10.1017/S0272263125101083 Published online by Cambridge University Press

2015) during which perceptual learning might be accelerated by integrating training earlier or later in language instruction, that is with intensive FI during the first year (Y1) of language study or with explicit phonetic instruction at a later stage. In the next section, we specifically address the typical format of university language degrees to further contextualize this research.

University language degrees

Degrees in language studies provide valuable opportunities to examine the development of L2 learners' perception and production of phonological features in the target language. Although the characteristics of the type of exposure that these programs involve are different from those of immersion in countries where the L2 is spoken, they typically involve substantial amounts of exposure to and interaction in the L2 over the course of three or four years. Moreover, in EFL contexts, English major programs often include dedicated courses in L2 pronunciation or phonetics, which provide learners with knowledge and practice in L2 phonological features.

Many studies exploring the effectiveness of HVPT have adopted the technique in university language programs involving extensive exposure and courses on phonetics and phonology (e.g., Aliaga-García, Mora & Cerviño-Povedano, 2011; Carlet & Cebrian, 2022; Carlet, Cebrian, Gavaldà & Gorba, 2022; Fouz-González & Mompean, 2021). However, whether the point at which training is implemented affects its effectiveness has not been explored in longitudinal studies that track learners' pronunciation development over time. Carlet et al. (2022) investigated the potentially facilitative role of metalinguistic knowledge in HVPT's effectiveness by comparing two studies: one in which they used metalinguistically experienced learners (in the second year [Y2] of their degree, with knowledge of English phonetics and phonology) and another one in which learners were metalinguistically naïve (in Y1 of their degree, before receiving any instruction in phonetics and phonology). The two studies were nearly identical in design, both involving HVPT with identification and discrimination tasks, although one included five HVPT sessions and a delayed posttest two months after training, whereas the other one included six HVPT sessions and a delayed posttest four months after training. The groups in the two studies improved their perception of English vowels after HVPT, especially after identification training. The gains made in nonword stimuli by the two groups following identification training in the two studies differed significantly, with learners who had prior phonetic training showing greater gains than those without such instruction. However, no significant differences were found in the gains for real word stimuli or between learners who received discrimination training.

Carlet et al.'s (2022) study offers valuable insights into how learners' metalinguistic awareness resulting from a phonetics and phonology course (PC) may enhance the effectiveness of HVPT. Nevertheless, whether the timing at which HVPT is implemented using the same HVPT set, within a single, longitudinally tracked cohort affects its effectiveness remains unexplored. Most experimental studies focus on relatively short-term gains using a pre-post–delayed design, providing limited insight into longer-term training effects and how the timing of training might regulate its effectiveness. However, placing instruction into a more comprehensive developmental window would allow a better understanding of its longitudinal effects (Nagle, 2021a).

https://doi.org/10.1017/S0272263125101083 Published online by Cambridge University Press

The current study

In this study, we compared the effectiveness of HVPT when aligned with the two WMOs, described in the Time and Timing in L2 Speech Learning section, namely the onset of intensive exposure (OIE) to the L2 through a university language degree and a phonetics and PC. On a conceptual level, aligning training with learners' OIE could catalyze perceptual learning by helping them perceive the L2 accurately before a period of concentrated exposure, leading to greater gains over time, gains that would be cemented in subsequent language courses. By the same token, introducing perceptual training into a PC could produce synergistic effects, insofar as explicit attention to pronunciation and pronunciation knowledge have been shown to facilitate L2 pronunciation learning (Carlet et al., 2022; Saito, 2013; 2019). Thus, the explicit phonetic information that learners receive in such a course could boost the effectiveness of the training (Uchihara et al., 2024).

Building on these considerations and on the impact that the timing of training may have on its effectiveness discussed in the preceding sections, in the current study we focused on the following research question: Does the effectiveness of HVPT vary depending on when it is implemented (at the OIE versus during a PC)?

To address this question, we followed a cohort of L2 English learners enrolled in an undergraduate English studies program over their first 2 years of university-level language instruction. In Y1, we administered 4 weeks of HVPT targeting a subset of particularly challenging English vowels to one group (G1) at the onset of their degree study and intensive exposure to the L2 (OIE) while the second group (G2) acted as control and received no training. In Y2, coinciding with all participants' enrollment in the PC module, G1 received no training while G2 received the same HVPT as G1 during Y1. We collected data before and after each group's training, yielding four data points (Figure 1). As is typical in perceptual training studies, we examined generalization by testing learners' performance on talkers and items that were not included in the training set.

We had no specific hypothesis regarding whether training during one window or the other would result in more robust learning. Instead, drawing upon existing literature, we reasoned that learners' vowel perception would improve as a result of HVPT. We also hypothesized that learners would show improvements outside of the HVPT windows, as a result of the general communicative language training they received.

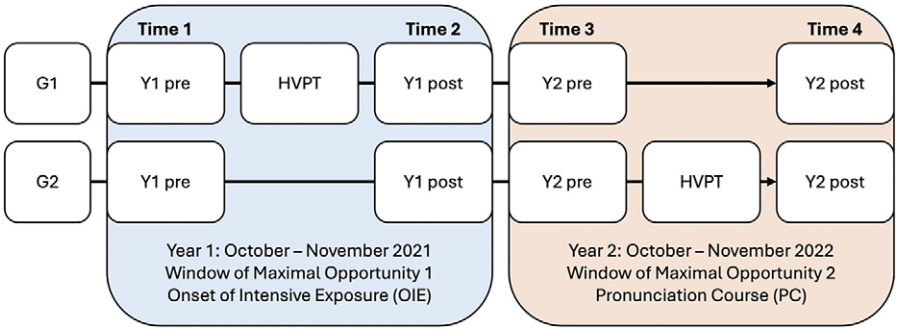

Figure 1. Study design with experimental groups (G1, G2), experimental phases (pretest, HVPT, posttest), timing (year 1, year 2), and hypothesized WMOs (OIE and PC).

https://doi.org/10.1017/S0272263125101083 Published online by Cambridge University Press

Method

Participants

Participants were recruited from a 4-year degree in English studies at a Spanish university. We tracked all participants over a 2-year window, randomly assigning them to two groups (G1, G2), as detailed above. G1 received training during the first semester of Y1 (2021) and G2 during the first semester of Y2 (2022). Of the 69 students who initially volunteered to take part in the study, only 36 (33 women, 3 men) completed the posttest in Y1 and were therefore eligible for inclusion in the study. Their mean age was 18.2 years (SD = 1.3). In Y2, 23 students took the pretest and 22 took the posttest (Table 1). Considering participant attrition, HVPT was delivered to 14 participants in G1 in Y1 and to 16 participants in G2 in Y2.

We aimed to recruit a relatively homogeneous group of participants in terms of language background to minimize variability in perceptual learning outcomes. All participants had English as their L2 with at least a B1 proficiency level according to the Common European Framework of Reference for Languages (CEFR), as determined by a university entrance exam. Moreover, at the onset of their degree program, they had undergone FI in EFL for over 10 years. All participants were predominantly monolingual Spanish speakers, although three participants deviated slightly from the monolingual profile: two balanced Spanish–German and Spanish–Arabic bilinguals and an L1 Italian speaker (with B2 proficiency in Spanish). None of the participants reported any auditory impairments. Further information about the participant's linguistic profile can be found in Appendix A.

The participants' degree involves ten six-credit courses per year (one credit = 10 hr of class) per semester, delivered in English. Participants were enrolled in a B2-level (CEFR) English language course at the time of training in Y1. Training in Y2 took place while students were enrolled in a C1-level English language course and in a compulsory English phonetics and PC providing them with a detailed account of the phonological system of Standard Southern British English (SSBE) as the main reference accent, a comparison with General American, connected speech processes, etc. (Mompeán González & Cutillas Espinosa, 2022 for an overview) as well as intensive transcription practice of written texts using SSBE (Mompean, 2017 for a description).

Materials

The same set of training stimuli as in Fouz-González and Mompean (2021) was used in this study. The HVPT paradigm was designed to help learners improve their perception of eight target SSBE vowels (/iː ɪ æ ʌ ɜː e ɒ ɔː/) that tend to be problematic for L1 Spanish learners of L2 English (e.g., Carlet & Cebrian, 2022; Fouz-González & Mompean, 2021; Monroy-Casas, 2001; Mora & Mora-Plaza, 2019) given orthographic interferences,

Table 1. Sample size by session/experimental phase

	Year 1 (OIE)			Year 2 (PC)		
	Time 1		Time 2	Time 3		Time 4
	pretest	HVPT	posttest	pretest	HVPT	posttest
G1	14	14	13	7	NA	6
G2	23	NA	23	16	16	16

https://doi.org/10.1017/S0272263125101083 Published online by Cambridge University Press

inaccurate perceptual assimilations, or acoustic differences between phonetically similar but not identical L1 and L2 vowels (Fouz-González & Mompean, 2021 for a detailed account).

Training and testing stimuli were recorded in a professional, soundproof studio by six native speakers of SSBE (three men, three women) who pronounced the stimuli in the carrier sentence *I say* ____. Four speakers were used for the training phase and the other two for the testing phase (split evenly between men and women). SSBE speakers were chosen (as opposed to speakers of other accentual varieties) to match the reference model used in instruction (see above).

The training stimuli consisted of 24 monosyllabic consonant-vowel-consonant (CVC) nonwords per target vowel (i.e., 192 nonwords). They were pronounced by four speakers and presented twice during training, thus totaling 1,536 stimuli. These nonwords were designed based on CV and VC phonotactic patterns in English. The choice of nonwords as opposed to real words was motivated by previous research suggesting that nonwords are more effective than real words for training L2 sounds, as they appear to facilitate greater attention to phonetic form by eliminating semantic interference and avoiding the activation of potentially inaccurate phonolexical representations (e.g., Ortega, Mora-Plaza & Mora, 2021; Thomson & Derwing, 2016). The nonwords featured the target vowels either preceded or followed by any of four fricative (/f v s z/) or two affricate (/ʧ ʤ/) consonants. There were two nonwords per target vowel starting with each of those consonants and two items ending in one of them. To provide further phonetic variability to the stimuli, some nonwords also started or ended with any of the six oral (/p b t d k g/) or two nasal (/m n/) stops. The sole non-obstruent consonant employed in nonwords was the approximant /j/ in onset position. This was essential to generate the required number of nonwords as intended, while ensuring real words were not inadvertently formed.

The testing stimuli were the same in all the pre- and posttests and included a subset of trained nonwords (12 per target vowel) from the training sessions as well as untrained nonwords (12 per target vowel) and a set of 32 (untrained) real words (for the stimuli, see Appendices B and C). The testing stimuli were produced by two different speakers, totaling 448 items. Untrained nonwords and real words were used to assess whether perceptual gains could be generalized to stimuli beyond trained nonwords. The untrained nonwords either started or ended with any of the same fricatives and affricates used during training (half the items in each case). Untrained real words featured the target sounds in four items per vowel.

Procedure

Considering that our goal was to evaluate whether the effectiveness of HVPT would vary depending on when it is implemented (at the OIE versus during a PC), the participants were randomly assigned to two experimental groups (G1 and G2) that received HVPT either in the Y1 or Y2 year of the degree. As mentioned above, G1 received training in Y1, coinciding with the OEI, and G2 received training in Y2, coinciding with the PC, a compulsory element of the degree program (see Figure 1 above).

Training and testing were conducted at a university language lab and administered with TP, an open-source application for speech perception tasks (Rato, Rauber, Kluge & Santos, 2015). Participants in both groups followed a 4-week HVPT program with a weekly identification task lasting approximately 30 minutes. During training,

https://doi.org/10.1017/S0272263125101083 Published online by Cambridge University Press

participants received immediate feedback on their responses after every trial, provided by TP with a tick or a cross indicating whether the participant's response was right or wrong. If the response was incorrect, TP showed the right answer, played the stimulus again, and the listener had to click on the right response before advancing to the next trial. Cumulative feedback was also provided at the end of each training session, showing the total score on the task (i.e., out of 384 trials), the number of incorrect responses, as well as the duration of the session.

The participants' perception and production[1] of the target sounds was assessed at four different points in time: before (T1) and after (T2) G1 received perceptual training in Y1 and before (T3) and after (T4) G2 was trained in Y2. The pretests took place a week before training and the posttests a week after it. All participants, regardless of group, took the pre- and posttests in both years of the study. The perception pretest and posttests were also identification tasks. Participants were shown the cumulative feedback screen at the end of the session, but they did not receive trial-by-trial feedback during testing.

The effectiveness of the exact same HVPT training was assessed in a previous study by the authors, with participants from the same degree, at the same university (Fouz-González & Mompean, 2021). In that study, training took place in Y2 of the degree, while students were enrolled in the compulsory PC described above. Statistical analyses revealed significant differences between the gains made by two experimental groups—differing only in one instructional variable unrelated to timing—and a control group who did not receive HVPT. Specifically, the two experimental groups showed larger improvements than the control group in trained nonwords and untrained nonwords, although no such advantage was observed for real words.

Evidence from (Fouz-González & Mompean, 2021) showed that this specific HVPT paradigm worked when administered in Y2 of the degree. However, since training in that study coincided with one of the potentially advantageous WMOs (the PC), it was necessary to evaluate whether this HVPT paradigm would also be effective if administered in Y1. Thus, before comparing the effectiveness of HVPT at the two WMOs of interest, we compared G1 and G2's perception gains between the Y1 pretest and the Y1 posttest (T1–T2, see the Results section).

Approach to analysis

We structured our analysis in two parts. First, to establish whether HVPT was effective, we compared G1 and G2 in Y1. For this analysis, G2 served as a control group for G1 because G2 participants did not receive training in Y1. Next, we compared both groups during Y2. By that time, both groups had received HVPT, but at different moments in the curriculum: G1 during Y1 and G2 during Y2. This analysis allowed us to determine if G2 improved significantly because of the HVPT they received, and crucially, if there were differences in achievement at the end of the 2-year period of observation. Since all participants were enrolled in a common curriculum consisting of intensive communicative language instruction in Y1 and Y2, and in pronunciation-related instruction in Y2, any observed differences at the end of Y2 could be attributed to the effect of HVPT

[1]Production was assessed with a picture naming task, a video-based narrative task, and an immediate imitation task. However, because the focus of the current study is on how the timing of HVPT affects perceptual learning, we do not describe the tests, the stimuli, or the production testing procedure in more detail, and we do not report the production data in this article.

https://doi.org/10.1017/S0272263125101083 Published online by Cambridge University Press

timing. In each of these analyses, we fit a model to the trained nonwords first, followed by the generalization conditions involving untrained nonwords and real words. For the generalization analysis, we set the trained nonwords as the baseline against which the two generalization conditions were compared to determine if there were significant differences in learning relative to the baseline trained nonwords. As a final step, we analyzed training trajectories for each group to examine whether the timing of HVPT affected the rate and shape of development during training.

We fit logistic mixed-effects models to the testing and training data using the lme4 package (version 1.1-35.3; Bates, Machler, Bolker & Walker, 2015) in R (version 4.4.0; R Core Team, 2024). Models included fixed effects of measurement point (pre, post HVPT, within years), and group (G1, HVPT in Y1; G2, HVPT in Y2) and their interaction. As a starting point, we included the following sensible set of random effects based on where we expected to observe interindividual variation in model-estimated parameters: by-participant random intercepts and slopes for measurement point; by-vowel random intercepts and slopes for measurement point, group, and their interaction; and by-word random intercepts. Only two speakers were used for testing, so we did not include speaker-level random effects due to the small number of units included in that grouping. The by-participant slopes for measurement point allowed us to capture variation in pre-post learning trajectories across participants. The by-vowel random slopes allowed us to quantify variation in pre-post trajectories across vowels as well as variability in the effect of group. For instance, the group who received HVPT (G1 in Y1, G2 in Y2) may have shown greater improvement on certain vowels over time compared to the other group. This variability can be quantified via the by-vowel random slopes for the interaction term.

We used the *buildmer* package (version 2.12; Voeten, 2023) to evaluate this random-effects structure, that is, to evaluate whether incorporating all the target random effects significantly improved model fit. In this way, we were able to strike a balance between conceptually motivated and data-driven random effects. In general, *buildmer* showed that the by-vowel random slopes for the Measurement Point × Group interaction could be dropped from the model. We fit all models using the bobyqa optimizer to reduce the likelihood of convergence issues. If we continued to encounter convergence issues or singular models, we simplified the random-effects structure until we achieved a converging model. We report model details for each analysis in the relevant section.

We decided to treat vowel targets as a random effect rather than a fixed effect for several reasons. Conceptually, we did not have any specific hypotheses related to how HVPT might have a differential effect according to the target vowel. We reasoned that vowels that participants initially identified more accurately might show a shallower learning trajectory because of their higher starting point, but that information is captured via the slope-intercept correlation in the random effect structure. Practically, given the modest sample size of the present study, trying to fit a Vowel × Measurement Point (× Group) interaction as a fixed effect would have created an underpowered and potentially uninterpretable model given that vowel has eight levels. More specifically, several significant by-vowel interactions would likely have arisen simply due to the estimation of many effects, but the credibility of those interactions and their practical significance would have been suspect. Thus, exploiting the random-effects structure to control for by-vowel variance in model-estimated parameters made sense given our research objectives.

After building the target models, we used the DHARMa package (version 0.4.6; Hartig, 2022) to simulate and inspect assumptions for model residuals; the sjPlot package (version 2.8.16; Lüdecke, 2024) to extract model estimates; and the ggeffects package (version 1.6.0; Lüdecke, 2018) to generate model-estimated predictions for plotting. For

https://doi.org/10.1017/S0272263125101083 Published online by Cambridge University Press

all models, we transformed log odds to odds ratios (*OR*s) and report *OR*s. An *OR* of 1.00 corresponds to a null effect, *OR* < 1.00 disfavors the outcome, and *OR* > 1.00 favors the outcome, relative to the baseline odds of responding correctly.

Results

First, we present descriptive statistics for each group at each measurement point. We then discuss the testing (pretest, posttest) results of G1 and G2 in Y1, in which G1 received training and G2 acted as control, followed by a similar analysis of G1 and G2 in Y2. After the within-year analyses, we analyze performance over both years, examining how each group's identification accuracy improved from one session to the next (including outside of training windows) and whether there were differences in overall accuracy between the two groups by the end of the study. We then turn to the training data, as we were also interested in whether the timing of training affected the rate and shape of learning over training sessions. Finally, we present an analysis of learners' performance as a function of vowel target, drawing upon the by-vowel random effects from the testing and training models.

Descriptive statistics and plots

As a first step, we computed descriptive statistics (means and *SD*s) for each group at each measurement point (Table 2). We also created a descriptive plot to visualize group and individual trajectories (Figure 2). Both groups had low overall accuracy at the beginning of the study but improved due to HVPT (G1 during Y1 and G2 during Y2). By the end of the study, both groups showed comparable levels of vowel identification accuracy. Yet, there was substantial individual variability, especially in terms of starting points for G2.

Year 1 analysis

For the Y1 analysis, we set G2 as the baseline (control condition) against which G1, the group who received training in Y1, was compared. The model fit to the trained nonwords is given in Table 3. The intercept refers to the pretest performance of the untrained group. The fact that the *OR* for that group did not reach significance and the confidence interval (CI) crossed zero suggests that G2 participants had approximately a 50% probability of responding correctly at pretest. Likewise, the non-significant G1 term shows that there was no difference between the two groups in the probability of responding correctly at pretest. The Y1 post term indicates how much G2 participants

Table 2. Means, *SD*s, and gain rates by training group and measurement point (Y1 pretest, Y1 posttest, Y2 pretest, Y2 posttest)

	G1		G2	
	M (*SD*)	Gain	*M* (*SD*)	Gain
Y1 pretest (T1)	.57 (.49)	—	.58 (.49)	—
Y1 posttest (T2)	.69 (.46)	.12	.61 (.49)	.03
Y2 pretest (T3)	.71 (.45)	.02	.69 (.46)	.07
Y2 posttest (T4)	.75 (.44)	.04	.77 (.42)	.08

Notes: Gain rates are expressed in decimal form. *M* = mean.

https://doi.org/10.1017/S0272263125101083 Published online by Cambridge University Press

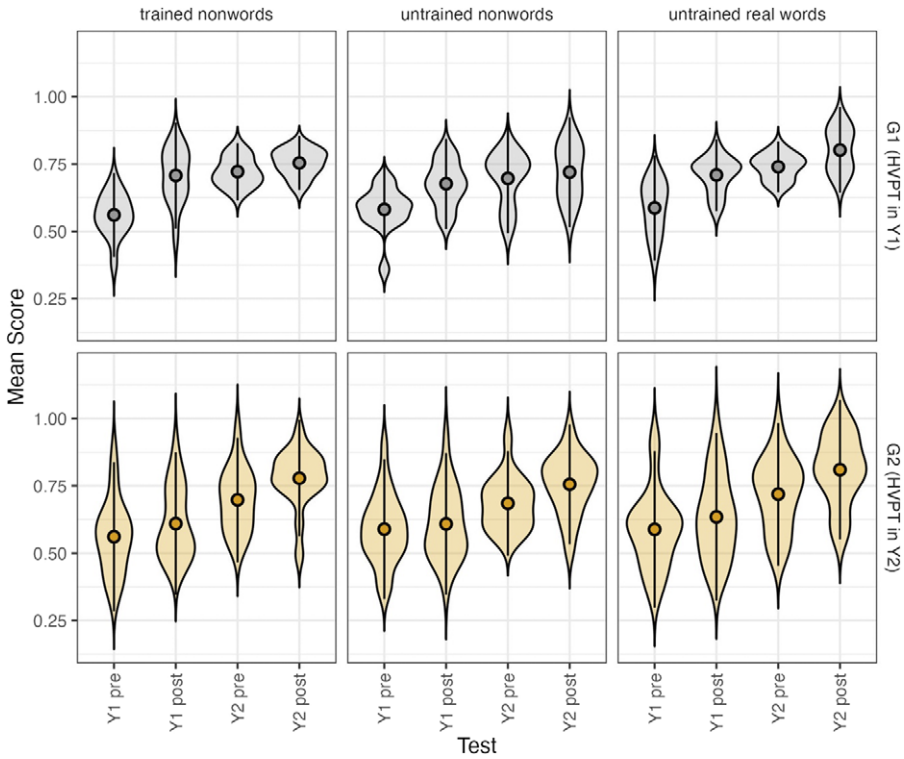

Figure 2. Descriptive identification performance across tests by group and word type (trained nonwords, untrained nonwords, untrained real words).
Note: The points represent the mean and the whiskers the *SD.*

Table 3. Summary of model fit to the trained nonwords (year 1)

	OR	SE	95% CI	z	p
Fixed effects					
Intercept	1.35	.27	[.91, 2.01]	1.50	.135
Y1 post	1.29	.12	[1.07, 1.56]	2.64	.008
G1	0.97	.22	[.63, 1.51]	–.12	.903
G1 × Y1 post	1.68	.19	[1.35, 2.09]	4.64	< .001
Random effects	*SD*	Correlation			
By-participant					
Intercepts	.57				
Slopes: Y1 post	.22	.06			
By-vowel					
Intercepts	.41				
Slopes: Y1 post	.20	.01			
Slopes: G1	.25	–.09	.40		
By-word					
Intercepts	.65				

Notes: Model syntax: glmer(Score ~ Group*Measurement + [1 + Measurement | Participant] + [1 + Measurement + Group | Vowel] + [1 | Word], data = data, family = "binomial", glmerControl[optimizer = "bobyqa"]). The baseline for Group was G2.

https://doi.org/10.1017/S0272263125101083 Published online by Cambridge University Press

improved at the posttest (T2). The significant $OR > 1$ demonstrates that the probability of responding correctly increased slightly, even for the group who did not receive HVPT. The significant interaction term, $OR > 1$, further indicates that G1 improved significantly more than G2. Put another way, G2 participants improved slightly, whereas G1 participants, who received HVPT, improved much more.

Next, we fit the generalization model incorporating a Group × Measurement Point × Word Type interaction into the model. For this analysis, we set the trained nonwords as the baseline for comparison. There were no statistically significant effects involving word type, which suggests that the gains observed for both groups (modest for G2, substantial for G1) were consistent across word types (full model reproducible using data and R code). Figure 3 plots model-estimated trajectories by group and word type.

Year 2 analysis

We repeated the analysis for the Y2 data. In this case, we set G1 as the baseline group because G2 received HVPT in Y2 (G1 completed their HVPT in Y1 but had no HVPT in Y2). In the model reported in Table 4, the intercept refers to G1 on the Y2 pretest (T3). According to the estimate, G1 participants were significantly above chance in terms of their probability of responding correctly, which is probably due to the fact that they received HVPT in Y1 and retained that learning into Y2. There were no significant differences between G1 and G2 on the Y2 pretest ($OR = .93$, $p = .824$). Although the estimate for G1 on the Y2 posttest (T4) suggested some improvement (Y2 post $OR = 1.26$, $p = .142$), that effect did not reach statistical significance. However, the interaction term did ($OR = 1.41$, $p = .015$), suggesting that G2 participants who received HVPT in Y2 improved significantly more from the Y2 pretest (T3) to the Y2 posttest (T4) than G1 participants did.

When we incorporated word type into the generalization model, no additional significant effects emerged, suggesting similar trajectories across the trained nonwords,

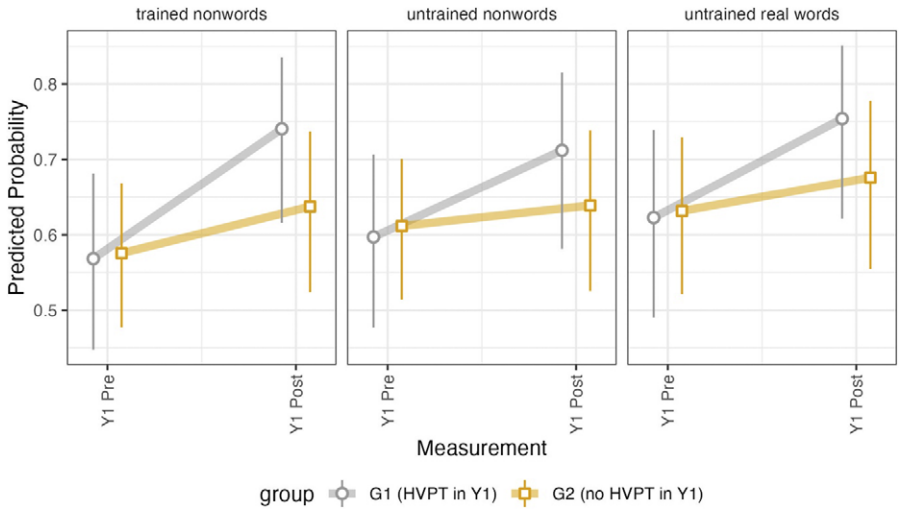

Figure 3. Model-estimated trajectories by training group and word type in year 1.

https://doi.org/10.1017/S0272263125101083 Published online by Cambridge University Press

Table 4. Summary of model fit to the trained nonwords (year 2)

	OR	SE	95% CI	z	p
Fixed effects					
Intercept	3.22	.95	[1.80, 5.76]	3.95	< .001
Y2 post	1.26	.20	[.93, 1.71]	1.47	.142
G2	.93	.29	[.51, 1.72]	–.22	.824
G2 × Y2 post	1.41	.20	[1.07, 1.85]	2.43	.015
Random effects	*SD*	Correlation			
By-participant					
Intercepts	.59				
Slopes: Y2 post	.14	.34			
By-vowel					
Intercepts	.45				
Slopes: Y2 post	.29	.38			
Slopes: G2	.38	–.26	.32		
By-word					
Intercepts	.87				

Notes: Model syntax: glmer(Score ~ Group*Measurement + [1 + Measurement | Participant] + [1 + Measurement + Group | Vowel] + [1 | Word], data = data, family = "binomial", glmerControl[optimizer = "bobyqa"]). The baseline for Group was G1.

untrained nonwords, and untrained real words. Model-estimated Y2 trajectories are shown in Figure 4.

Analysis of both years

In addition to examining the effect of HVPT within each year (Y1 for G1, Y2 for G2), we were also interested in (a) how each group's identification accuracy improved from one session to the next, including outside of training windows, which would provide evidence for the type of automatic development that can sometimes happen in the absence of targeted training (Derwing, 2010); and (b) whether, by the end of the study, there were significant differences between the two groups in terms of their overall accuracy.

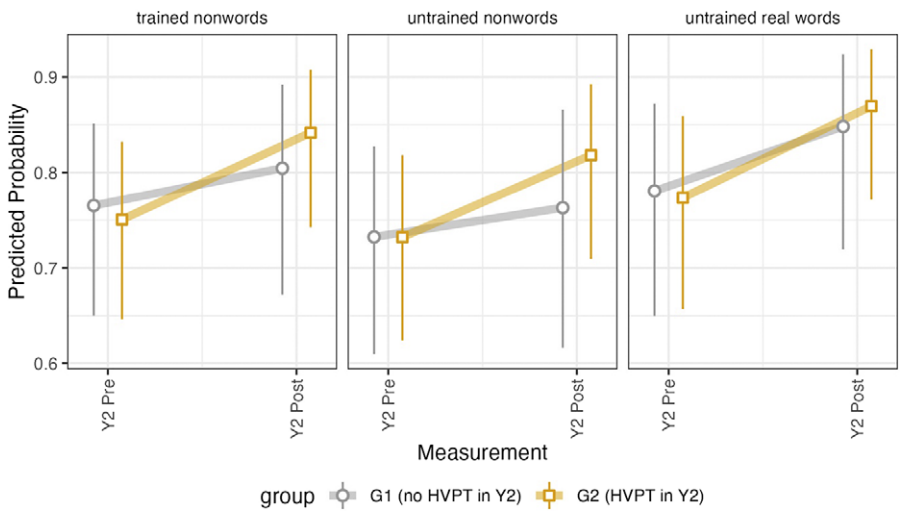

Figure 4. Model-estimated trajectories by training group and word type in year 2.

https://doi.org/10.1017/S0272263125101083 Published online by Cambridge University Press

Table 5. Planned time-wise comparisons within groups

	G1 (HVPT in Y1)			G2 (HVPT in Y2)		
Comparison	OR	SE	p	OR	SE	p
Y1 pre (T1) / Y1 post (T2)	1.98	.40	.003	1.20	.07	.004
Y1 post (T2) / Y2 pre (T3)	1.14	.20	.863	1.52	.22	.018
Y2 pre (T3) / Y2 post (T4)	1.34	.17	.062	1.78	.24	< .001
Y1 pre (T1) / Y2 post (T4)	3.03	.57	< .001	3.24	.55	< .001
Y1 post (T2) / Y2 post (T4)	1.53	.26	.042	–	–	–
Y1 pre (T1) / Y2 pre (T2)	–	–	–	1.82	.28	< .001

Note: The multivariate *t* distribution was used to adjust *p* values for multiple comparisons (5) for each group.

To address (a), we fit separate models to each group, using the *emmeans* package (version 1.11.0; Lenth, 2024) to compare consecutive data points, pretest and posttest outside of training (Y2 for G1, Y1 for G2), and Y1 pretest (T1) to Y2 posttest (T4), as an overall measure of how much each group improved. Given that previous analyses did not provide compelling evidence for significant differences by word type, we report pooled estimates here, averaging over trained nonwords, untrained nonwords, and untrained real words. As shown in Table 5, G1 improved significantly during Y1, which was expected because that group received HVPT in that year. There was no significant gain from the Y1 posttest (T2) to the Y2 pretest (T3). There was evidence of some improvement during Y2, although the Y2 pre-post comparison did not reach statistical significance. Considering the entire non-training window, from the Y1 posttest (T2) to the Y2 posttest (T4), there was evidence of statistically significant improvement. Analyses for G2 showed statistically significant improvement across all comparisons, including during Y1, when that group did not receive HVPT. It seems, then, that the intensive communicative English instruction that learners received during Y1 may have been enough to catalyze development even for G2 learners who did not receive HVPT until Y2, whereas learning for G1 participants, who received HVPT in Y1 but not Y2, was not as marked during Y2. Nonetheless, both groups appeared to reach the same endpoint, achieving comparable levels of identification accuracy. To make this analysis as interpretable as possible, we have input the *OR* estimates into the original study design figure (Figure 5).

To examine differences in final performance, we fit a model to only the Y2 posttest (T4) data with fixed effects of Group, Word Type, and Group × Word Type to determine if there were significant differences between groups in terms of their performance at the end of the study. This model revealed no significant differences across the board, which suggests that both G1, who had received HVPT in Y1, and G2, who had received HVPT in Y2, achieved similar levels of identification accuracy by the end of the 2-year period of observation. To visualize trajectories over the full 2-year period, model-estimated trajectories are plotted in Figure 6.

Training data

The results of the testing data analysis showed that both groups improved significantly when they received HVPT: G1 in Y1 and G2 in Y2. By the end of the study, both groups were significantly more likely to identify vowels accurately, and there were no significant differences between groups, suggesting that the timing of HVPT did not result in differential pre-post gains. We were also interested in examining whether the timing of training affected the rate and shape of learning over training sessions. To that end, we

https://doi.org/10.1017/S0272263125101083 Published online by Cambridge University Press

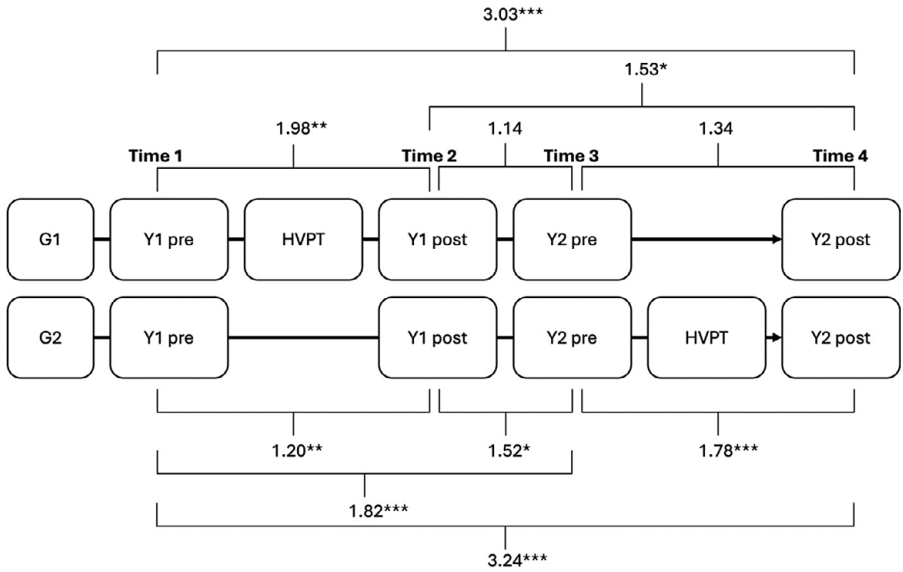

Figure 5. Summary of time-wise comparisons within groups.

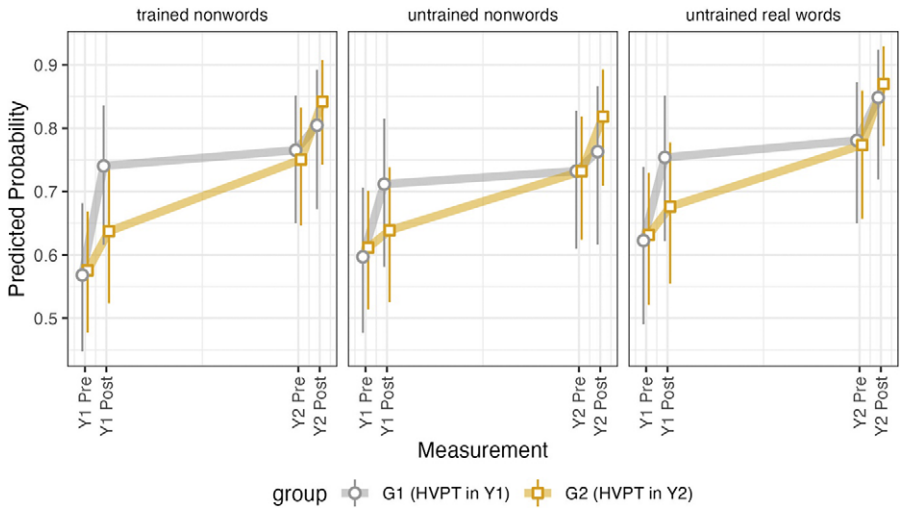

Figure 6. Combined model-estimated trajectories by training group and word type.

analyzed the four sessions of training data, incorporating session[2], group, and their interaction as fixed effects. We set G1 as the baseline condition against which G2 was compared. We included by-participant random intercepts and slopes for session, by-vowel random intercepts and slopes for session and group, and by-word random

[2]Plotting of the observed data showed that the training trajectories were mostly linear, which suggests that higher order polynomial terms used to estimate curvature were not warranted. We therefore limited our analysis to the linear effect of session.

https://doi.org/10.1017/S0272263125101083 Published online by Cambridge University Press

intercepts, to make this model comparable to the models we fit to the testing data. Because four speakers were included in the training stimulus set, we attempted to fit by-speaker random intercepts and slopes to control for speaker-level variance in model-estimated parameters. However, models were either singular or failed to converge, so we dropped the by-speaker random effects. The training model therefore had the same random effects structure as the testing models.

As reported in Table 6, G1 had a slightly better than average chance of responding correctly to a training trial at the outset of training (Intercept OR = 1.60, p = .012). Moreover, the statistically significant effect for G2 (OR = 1.59, p = .009) indicates that G2 had a higher likelihood of responding correctly at the outset of training. This result makes sense because G2 trained in Y2, by which time they had a higher starting point. Both groups improved significantly across HVPT sessions, as indicated by the statistically significant effect of session (OR = 1.17, p < .001). However, the non-significant interaction (OR = .96, p = .202) suggests that the groups were not significantly different in terms of their rate of learning. Figure 7 plots model-estimated training trajectories for both groups.

By-vowel random effects

As previously described, we treated vowel as a random effect grouping, fitting by-vowel random slopes for measurement point (testing data) and session (training data). To gain insight into how performance and improvement varied as a function of vowel target, we extracted and plotted the by-vowel random intercepts and slopes. Figure 8 shows the plot for the testing data. The first thing to observe is that individual vowels showed the same pattern as the overall, combined trajectory. For G1, who were trained in Y1, all vowels improved from pretest to posttest, but the magnitude of the gain varied, as expected. The same pattern is evident for G2. For both groups, several vowels showed limited improvement, especially from pre- to posttest in Y2 (e.g., /iː ɪ ɜː ɔː/). The /ʌ/ vowel proved especially challenging for both groups, showing some improvement but a much lower overall level of identification accuracy, compared to the other vowel targets.

Table 6. Summary of model fit to the training data

	OR	SE	95% CI	z	p
Fixed effects					
Intercept	1.60	.30	[1.11, 2.30]	2.52	.012
Session	1.17	.04	[1.09, 1.26]	4.43	< .001
G2	1.59	.28	[1.12, 2.25]	2.61	.009
G2 × Session	.96	.03	[.90, 1.02]	−1.28	.202
Random effects	*SD*	Correlation			
By-participant					
Intercepts	.43				
Slopes: Session	.06	.33			
By-vowel					
Intercepts	.37				
Slopes: Session	.08	−.19			
Slopes: G2	.13	.11	.94		
By-word					
Intercepts	.63				

Notes: Model syntax: glmer(Score ~ Group*Session + [1 + Session | Participant] + [1 + Group + Session | Vowel] + [1 | Word], data = data, family = "binomial", glmerControl[optimizer = "bobyqa"]). G1 was the reference level for Group.

https://doi.org/10.1017/S0272263125101083 Published online by Cambridge University Press

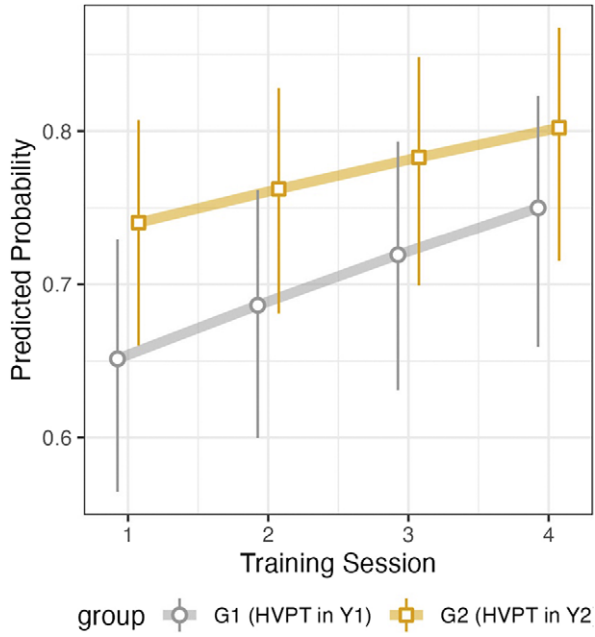

Figure 7. Model-estimated training trajectories by training group.

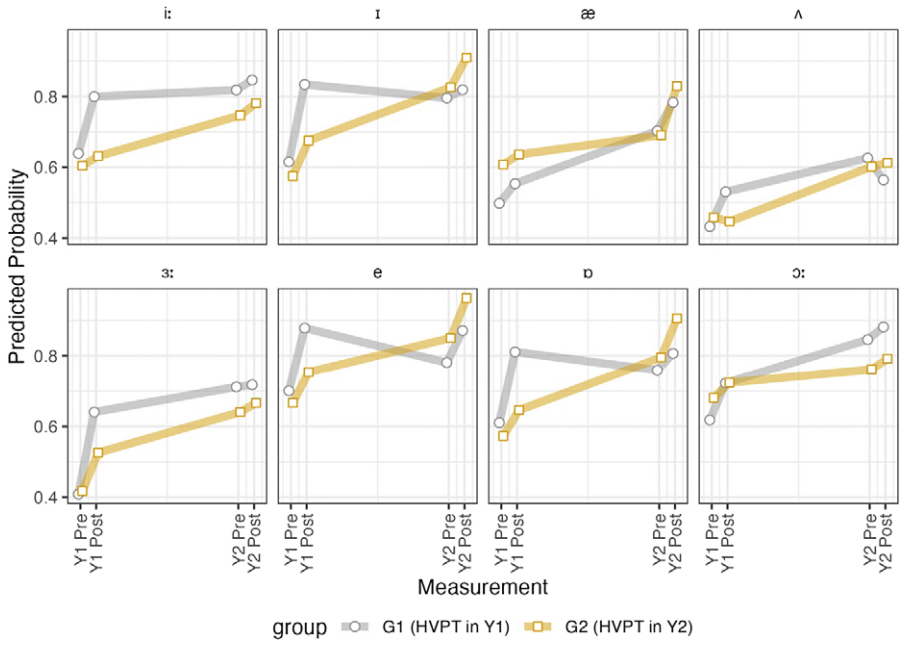

Figure 8. Model-estimated testing trajectories by vowel based on by-vowel random effects.

https://doi.org/10.1017/S0272263125101083 Published online by Cambridge University Press

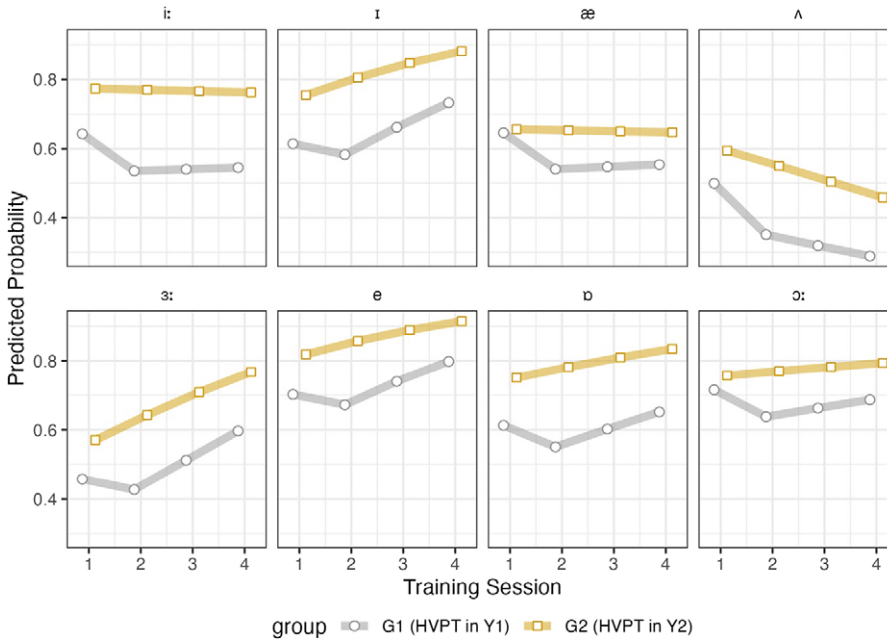

Figure 9. Model-estimated training trajectories by vowel based on by-vowel random effects.

The by-vowel analysis also suggests that some vowels were subject to backsliding for G1 after training (in the periods between T2 and T3 or between T3 and T4), but performance on the Y2 tests for G1 remained above their Y1 posttest performance.

We also plotted the by-vowel random effects for the training data following the same procedure. As shown in Figure 9 performance on /ɪ ɜː e ɒ ɔː/ improved to varying degrees, whereas performance on /iː æ ʌ/ was either relatively flat or appeared to decline over time. Considering both types of data, one interesting finding is that performance on /ʌ/ improved slightly between testing times, whereas the training data showed the opposite trend. Crucially, decreasing performance during training does not necessarily indicate worsening ability, but rather could be a sign of the destabilization and reorganization of L2 categories (Best & Tyler, 2007), which in terms of testing could lead to better overall performance.

The by-vowel data could be indicative of meaningful subpatterns that could be explored in future studies (i.e., grouping vowels by initial assimilation patterns, pretest difficulty level, etc., and including that conceptually motivated variable in interaction with time). However, it is also the case that some by-vowel variability in starting points and trajectories is to be expected, in the same way that similar variability is to be expected across participants, items, and so on. Thus, we caution against the over-interpretation of this data.

Discussion

This study investigated whether the effectiveness of HVPT would vary depending on when it is implemented (at the OIE versus during a PC). We hypothesized the existence of two potential WMOs during learners' degree program that might enhance the effectiveness of HVPT: WMO1, coinciding with the onset of learners' intensive exposure (OIE) to the L2 as part of a university language degree that provides learners

https://doi.org/10.1017/S0272263125101083 Published online by Cambridge University Press

with extensive communicative L2 instruction; and WMO2, coinciding with learners' enrollment in an L2 pronunciation-related English phonetics and PC providing them with explicit information on the L2 phonology.

Given these two WMOs, G1 improved by 12% from the Y1 pretest (T1) to the Y1 posttest (T2), significantly outperforming G2, which served as the control group during Y1. This provides evidence that the HVPT implemented in this study was effective at promoting perceptual learning. During Y2, G2 improved by 8% from the Y2 pretest (T3) to the Y2 posttest (T4), a gain that was significantly greater than the one made by G1, who did not receive HVPT during that year. Although within each year the group receiving HVPT showed significantly more gains than the group who did not, by the end of the study, both groups reached a comparable level of identification accuracy, suggesting that the timing of HVPT did not have a strong impact on the ultimate amount of learning achieved at the end of the 2-year window of observation. Additionally, our analysis of the training data also shows that both groups learned at approximately the same rate, reinforcing the view that timing did not play a significant role in the present study.

Our findings have implications on our understanding of the longitudinal development of perceptual accuracy and the role that training plays in that development. The results obtained indicate that the beginning of an extensive exposure in a foreign language learning environment represented by L2 learners' enrollment in a major in English is a WMO that has a positive impact on the development of learners' perception of L2 sounds. To that point, G2 showed improvement before HVPT as shown in the Y1 posttest (T2) and the Y2 pretest (T3). A similar pattern was observed in G1, who improved in Y1 due to HVPT but also continued to show growth after it, as shown in the Y2 pretest (T3) and posttest (T4). Thus, our findings add to the growing body of evidence suggesting that periods of intensive L2 exposure are beneficial for perceptual learning, not only in naturalistic settings (Derwing & Munro, 2015) but also in instructed contexts (this study).

Our results also have implications, albeit limited, for our understanding of the impact of PCs on perceptual learning. By the end of two 2-year period of observation, both groups had reached a comparable level of perceptual accuracy in identifying challenging L2 English vowel contrasts. In our study, combining HVPT with a PC was no more effective than combining it with the OIE. Nevertheless, the evidence available suggests that the metalinguistic knowledge obtained through this type of course may enhance the effectiveness of perceptual training. In Carlet et al.'s (2022) study, for example, Y2 English majors enrolled in an English phonetics course made larger gains after HVPT than first-year learners without such metalinguistic knowledge. It is important to note, though, that their Y2 learners had also had the benefit of more exposure to the L2. The WMO-like character of L2 pronunciation-related courses, therefore, requires further exploration, including assessing the possible impact of OIE and PCs independently from a longitudinal perspective.

Apart from the contribution of the onset and continuation of extensive exposure to the L2 to learners' development of L2 pronunciation, the current study has also shown that HVPT led to perceptual gains that were robust in both experimental groups, adding to the substantial body of evidence on the effectiveness of this technique (e.g., Thomson, 2018; Uchihara et al., 2024). It is especially noteworthy that it took G2 a whole year of intensive language instruction to make the same progress that G1 made with just four HVPT sessions of approximately 30 min. The effectiveness of HVPT can be considered independent of learners' engagement in intensive L2 exposure and coursework provided by their degree program because G1 significantly outperformed

https://doi.org/10.1017/S0272263125101083 Published online by Cambridge University Press

G2 on the Y1 posttest (with G2 serving as a control group to assess the effectiveness of HVPT in Y1).

It is also important to note that the testing stimuli were pronounced by two speakers that were not featured during training, which means that learning generalized to untrained voices. Also, and in line with the results of previous studies (e.g., Carlet et al., 2022; Fouz-González & Mompean, 2021), the data show that gains occurred consistently in the three types of testing stimuli used (trained nonwords, untrained nonwords, untrained real words), which offers further evidence of generalization to untrained stimuli.

Finally, the by-vowel analysis showed that even if training helped learners improve their perception of all the target vowels, the magnitude of the gains varied. The gains for some vowels were relatively modest and there was even evidence of backsliding by G1 participants for several vowels a year after training. While G1's performance on the Y2 posttest (T4) remained above their Y1 posttest levels (T2) for most vowels, these findings suggest that a single training period (or a period of such a short length) may not be sufficient for every target segment. Certain segments may require follow-up training sessions to ensure lasting gains. Some by-vowel variation is expected because, on the one hand, some vowels may be easier to perceive (or at least, easier to perceive for some learners), and those vowels may show a higher starting point, which could then affect the rate of improvement. The small but negative slope-intercept correlation in the training model suggests that this was the case.

Recommendations for L2 pronunciation instruction

The results of this study allow us to make some tentative recommendations for the field of L2 pronunciation. First, language practitioners should take heart from the fact that HVPT was effective at both time points examined in this study, suggesting that what matters most is that it is implemented, rather than precisely when. Learners may benefit from training in the early stages of a language program, which may help them process subsequent input more efficiently, or later in the program, capitalizing on a longer period of previous exposure to the L2 and starting training perhaps with more precision. Such flexibility is likely to be beneficial for teachers who may find it difficult to implement HVPT at specific moments during an instructional program. Similarly, even if teachers are not able to implement HVPT, it seems probable that learners' perception will improve somewhat through intensive language instruction, although it is also important to note that the typical L2 learner may not have the same type or amount of exposure to the target language as did the English majors in this study. Of course, it seems clear that the best course of action is combining all the elements in this research: intensive language training, HVPT, and explicit information on pronunciation. When combined, these elements may act synergistically to help learners shore up gaps in their perception, channeling all learners, regardless of starting point, toward a positive outcome.

Limitations and future directions

The current study has several limitations, which also offer directions for future research. First, while most participants were predominantly monolingual, three bilinguals were included in the sample. While bilingual individuals may differ in perceptual learning abilities compared to monolinguals, the small number of non-monolingual participants is unlikely to have significantly influenced the overall

https://doi.org/10.1017/S0272263125101083 Published online by Cambridge University Press

results. However, this is an important variable and should be controlled for as much as possible in future studies.

Another point worth mentioning is the specific participant profile in the current study (i.e., university English majors). This profile allowed us to study a WMO such as the period during which participants were enrolled in an English phonetics and PC. However, this is not the reality for many learners of L2 English or, indeed, other instructed L2 learners who do not have the chance to enroll in a PC in their program of study. Therefore, studies such as the above could be conducted with learners with either less exposure to English or at different WMOs than the ones studied above. Future longitudinal studies could also be carried out in an FL context with HVPT but without PC, as well as in FL contexts without HVPT. This could include younger learners in pre-university education where other WMOs may be identified, such as their participation in courses or educational programs that boost their immersive interaction and exposure to the L2 (e.g., Content and Language Integrated Learning programs). For this reason, it is also important to carefully consider the generalizability of the present findings. Most proximally, we can have confidence that they are likely to generalize to learners enrolled in similar instructional programs where intensive communicative language training is coupled with at least one course in pronunciation. Whether they generalize to other contexts where learners do not take courses in pronunciation or receive pronunciation instruction as part of general language instruction is an open question. We believe that findings would hold, but perhaps the magnitude of the gain would vary depending on context-specific timing and instructional characteristics.

We also chose not to incorporate what could be considered a true control group, a group who received the same general instruction and took the same compulsory PC course as the HVPT groups but did not receive HVPT at any point. It is possible that general instruction and PC without HVPT could lead to comparable gains. At the same time, we believe that such an outcome is unlikely given the positive within-year effect of HVPT for the groups included in this study as well as accumulated findings demonstrating the efficacy of HVPT in general and for this learner population in particular. What seems certain is that HVPT can accelerate learning. Thus, even if learners can reach a similar level of perceptual accuracy without HVPT, it will likely take them much longer to do so.

The current study also suffers from a recurrent problem in longitudinal data analysis, that is, the substantial number of dropouts. Unfortunately, there is no easy solution for this problem, making replication studies a particularly valuable means of corroborating and adding additional insight to longitudinal experimental studies. Thus, the findings presented here, while promising, should be regarded as preliminary pending future studies, including replication studies.

Finally, in this study, we opted to treat vowel as a random effect in our models because we aimed to estimate the mean, pooled trajectory, considering all vowels, as well as vowel-specific deviance around that mean trajectory. In future work, vowel could be treated as a fixed effect provided researchers have hypotheses related to why certain types of vowels might show distinct developmental trajectories and the sample size to estimate such vowel-specific trajectories.

Conclusion

The current study was designed to examine whether the moment at which HVPT is administered has an impact on its effectiveness. The study involved two academic years

https://doi.org/10.1017/S0272263125101083 Published online by Cambridge University Press

with one group of learners receiving HVPT in the first semester of Y1 and the other in the first semester of Y2. Findings highlight the benefits of HVPT regardless of when it is incorporated in the curriculum. Overall, we believe it is crucial to evaluate training as an element of a comprehensive approach to language instruction, considering the entire timeline of instructional programs.

Data availability statement. All research materials, data, and analysis code are available at https://osf.io/r4uxh/.

Acknowledgments. Charlie Nagle gratefully acknowledges financial support for this project by the Fulbright U.S. Scholar Program, which is sponsored by the U.S. Department of State and the Spain-U.S. Fulbright Commission. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Fulbright Program, the Government of the United States, or the Spain-U.S. Fulbright Commission. The study is also funded by grant PID2022-137876NB I00, by MCIN/AEI/10.13039/501100011033, and by the ESF+.

Contributions. Author roles were classified using the Contributor Role Taxonomy (CrediT; https://credit.niso.org/) as follows: Charlie Nagle: conceptualization, data curation, formal analysis, methodology, writing—original draft, and writing—review and editing; Jose Mompean: conceptualization, data curation, investigation, methodology, writing—original draft, and writing—review and editing; and Jonás Fouz-González: conceptualization, data curation, investigation, methodology, writing—original draft, writing—review and editing.

Competing interests. The authors declare none.

References

Aliaga-García, C., Mora, J. C., & Cerviño-Povedano, E. (2011). L2 speech learning in adulthood and phonological short-term memory. *Poznań Studies in Contemporary Linguistics*, *47*, 1–14. https://doi.org/10.2478/psicl-2011-0002

Bates, D., Machler, M., Bolker, B., & Walker, S. (2015) Fitting linear mixed-effects models using *lme4*. *Journal of Statistical Software*, *67*, 1–48.

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). John Benjamins.

Brekelmans, G., Lavan, N., Saito, H., Clayards, M., & Wonnacott, E. (2022). Does high variability training improve the learning of non-native phoneme contrasts over low variability training? A replication. *Journal of Memory and Language*, *126*, 104352. https://doi.org/10.1016/j.jml.2022.104352

Carlet, A., & Cebrian, J. (2022). The roles of task, segment type, and attention in L2 perceptual training. *Applied Psycholinguistics*, *43*, 271–299. https://doi.org/10.1017/S0142716421000515

Carlet, A., Cebrian, J., Gavaldà, N., & Gorba, C. (2022). Does metalinguistic knowledge about the L2 enhance the effectiveness of L2 perceptual training? In B. Blecua, J. Cicres, M. Espejel, & M. J. Machuca (Eds.), *Propuestas en fonética experimental: Enfoques metodológicos y nuevas tecnologías* (pp. 31–35). Universitat de Girona-Servei de Publicacions.

Casillas, J. V. (2020). The longitudinal development of fine-phonetic detail: Stop production in a domestic immersion program. *Language Learning*, *70*, 768–806. https://doi.org/10.1111/lang.12392

Cebrian, J., Gavaldà, N., Gorba, C., & Carlet, A. (2024). Differential effects of identification and discrimination training tasks on L2 vowel identification and discrimination. *Studies in Second Language Acquisition*, *46*, 1069–1093. https://doi.org/10.1017/S0272263124000408

Derwing, T. M. (2010). Utopian goals for pronunciation teaching. In J. Levis & K. LeVelle (Eds.), *Proceedings of the 1st Pronunciation in Second Language Learning and Teaching Conference*, Iowa State University (pp. 24–37). Iowa State University.

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins.

Flege, J. E., & Bohn, O-S. (2021). The revised speech learning model (SLM-r). In R. Wayland (Ed.), *Second language speech learning: Theoretical and empirical progress* (pp. 3–83). Cambridge University Press.

Fouz-González, J., & Mompean, J. A. (2021). Exploring the potential of phonetic symbols and keywords as labels for perceptual training. *Studies in Second Language Acquisition*, *43*, 297–328. https://doi.org/10.1017/S0272263120000455

Hartig, F. (2022). *DHARMa: Residual diagnostics for hierarchical (multi-level / mixed) regression models.* http://florianhartig.github.io/DHARMa/

Kim, D., Clayards, M., & Goad, H. (2018). A longitudinal study of individual differences in the acquisition of new vowel contrasts. *Journal of Phonetics*, *67*, 1–20. https://doi.org/10.1016/j.wocn.2017.11.003

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2007). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*, 979–1000.

Lenth, R. V. (2024). *emmeans: Estimated marginal means, aka least-squares means* (R package version 1.10.0). https://rvlenth.github.io/emmeans/

Lively, S. E., Logan, J. S., & Pisoni, D. B. (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*, 1242–1255. https://doi.org/10.1121/1.408177

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y. I., & Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /1/ III: Long-term retention of new phonetic categories. *Journal of the Acoustical Society of America*, *96*, 2076–2087. https://doi.org/10.1121/1.410149

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*, 874–886. https://doi.org/10.1121/1.1894649

Lord, G. (2010). The combined effects of immersion and instruction on second language pronunciation. *Foreign Language Annals*, *43*, 488–503. https://doi.org/10.1111/j.1944-9720.2010.01094.x

Lüdecke, D. (2018). *ggeffects*: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, *3*, 772, https://doi.org/10.21105/joss.00772

Lüdecke, D. (2024). *sjPlot: Data visualization for statistics in social science* (R package version 2.8.16). https://CRAN.R-project.org/package=sjPlot.

Mompean, J. A. (2017). Doing phonetic transcription in a modern language degree. In J. A. Cutillas Espinosa, J. M. Hernandez Campoy, R. M. Manchon Ruiz, & F. Mena Martinez (Eds.), *Estudios de filologia inglesa: Homenaje al profesor Rafael Monroy* (pp. 479–505). EditUM.

Mompeán González, J. A., & Cutillas Espinosa, J. A. (2022). Fonetica y fonologia inglesas [English Phonetics and Phonology]. EditUM. https://doi.org/10.6018/editum.2958

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.

Mora, J. C. (2014). The role of onset level on L2 perceptual phonological development after formal instruction and study abroad. In Pérez-Vidal, C. (Ed.). (2014). *Language acquisition in study abroad and formal instruction contexts* (pp. 167–194). John Benjamins.

Mora, J. C., & Mora-Plaza, I. (2019). Contributions of cognitive attention control to L2 speech learning. In A. M. Nyvad, M. Hejná, A. Højen, A.B. Jespersen & M. Hjortshøj Sørensen (Eds.), *A sound approach to language matters: In honor of Ocke-Schwen Bohn* (pp. 477–499). Aarhus University.

Mora, J., Ortega, M., Mora-Plaza, I., & Aliaga-García, C. (2022). Training the pronunciation of L2 vowels under different conditions: The use of non-lexical materials and masking noise. *Phonetica*, *79*, 1–43. https://doi.org/10.1515/phon-2022-2018

Nagle, C. (2018). Examining the temporal structure of the perception-production link in second language acquisition: A longitudinal study. *Language Learning*, *68*, 234–270. https://doi.org/10.1111/lang.12275

Nagle, C. (2021a). Assessing the state of the art in longitudinal L2 pronunciation research: Trends and future directions. *Journal of Second Language Pronunciation*, *7*, 154–182. https://doi.org/10.1075/jslp.20059.nag

Nagle, C. (2021b). Revisiting perception–production relationships: Exploring a new approach to investigate perception as a time-varying predictor. *Language Learning*, *71*, 243–279. https://doi.org/10.1111/lang.12431

Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics*, *25*, 26–45. https://doi.org/10.1017/S0267190505000024

Ortega, M., Mora-Plaza, I., & Mora, J. C. (2021). Differential effects of lexical and non-lexical HVPT on the production of L2 vowels. In A. Kirkova-Naskova, A. Henderson, & J. Fouz-González (Eds.), *English pronunciation instruction: Research-based insights* (pp. 327–355). John Benjamins.

Pérez-Vidal, C. (Ed.). (2014). *Language acquisition in study abroad and formal instruction contexts*. John Benjamins.

Perrachione, T. K., Lee, J., Ha, L. Y., & Wong, P. C. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America*, *130*, 461–472. https://doi.org/10.1121/1.3593366

R Core Team (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rato, A. (2014). Effects of perceptual training on the identification of English vowels by native speakers of European Portuguese. *Proceedings of the International Symposium on the Acquisition of Second Language Speech – Concordia Working Papers in Applied Linguistics*, *5*, 529–546.

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). Palgrave Macmillan.

Saito, K. (2013). Reexamining effects of form-focused instruction on L2 pronunciation development: The role of explicit phonetic information. *Studies in Second Language Acquisition*, *35*, 1–29. https://doi.org/10.1017/S0272263112000666

Saito, K. (2019). Individual differences in second language speech learning in classroom settings: Roles of awareness in the longitudinal development of Japanese learners' English /ɹ/ pronunciation. *Second Language Research*, *35*, 149–172. https://doi.org/10.1177/0267658318768342

Sakai, M., & Moorman, C. (2018). Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research. *Applied Psycholinguistics*, *39*, 187–224. https://doi.org/10.1017/S0142716417000418

Silpachai, A. (2020). The role of talker variability in the perceptual learning of Mandarin tones by American English listeners. *Journal of Second Language Pronunciation*, *6*, 209–235. https://doi.org/10.1075/jslp.19010.sil

Thomson, R. I. (2012). Improving L2 listeners' perception of English vowels: A computer-mediated approach. *Language Learning*, *62*, 1231–1258. https://doi.org/10.1111/j.1467-9922.2012.00724.x

Thomson, R. I. (2018). High Variability [Pronunciation] Training (HVPT): A proven technique about which every language teacher and learner ought to know. *Journal of Second Language Pronunciation*, *4*, 208–231. https://doi.org/10.1075/jslp.17038.tho

Thomson, R. I., & Derwing, T. M. (2016). Is phonemic training using nonsense or real words more effective? In J. Levis, H. Le., I. Lucic, E. Simpson, & S. Vo (Eds.), *Proceedings of the 7th Pronunciation in Second Language Learning and Teaching Conference* (pp. 88–97). Iowa State University.

Tremblay, A., Kim, H., Kim, S., & Cho, T. (2023). Perceptual training enhances the use of vowel quality cues to lexical stress: The benefits of intonational variability. In R. Skarnitzl, & J. Volín (Eds.) *Proceedings of the 20th International Congress of Phonetic Sciences – ICPhS 2023* (pp. 2111–215). International Phonetic Association.

Uchihara, T., Karas, M., & Thomson, R. I. (2024). Does perceptual high variability phonetic training improve L2 speech production? A meta-analysis of perception-production connection. *Applied Psycholinguistics*, *45*, 591–623. https://doi.org/10.1017/S0142716424000195

van Leussen, J-W., & Escudero, P. (2015) Learning to perceive and recognize a second language: The L2LP model revised. *Frontiers in Psychology*, *6*, 1000. https://doi.org/10.3389/fpsyg.2015.01000

Voeten, C. C. (2023). *buildmer: Stepwise elimination and term reordering for mixed-effects regression*. https://cran.r-project.org/web/packages/buildmer/index.html

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America*, *106*, 3649–3658. https://doi.org/10.1121/1.428217

Zhang, X., Cheng, B., & Zhang, Y. (2021). The role of talker variability in nonnative phonetic learning: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research*, *64*, 4802–4825. https://doi.org/10.1044/2021_jslhr-21-00181

Appendix A. Detailed summary of participant characteristics (*N* = 35)

Characteristics	*M*	*SD*	Range
Age (years)	18.20	1.28	17–23
Age of onset L2 English (years)	6.31	3.07	3–14
Years L2 English study	13.00	2.39	7–16
Global English proficiency[b]	2.55	.57	1.50–3.50
English listening proficiency[a]	2.77	.73	2–4
English speaking proficiency[a]	2.46	.78	1–4
English reading proficiency[a]	2.60	.60	2–4
English writing proficiency[a]	2.38	.55	1–3
Study abroad length (months)[c]	1.78	2.13	.25–7
Study abroad experience in English context	Yes: 10	No: 25	
Global CEFR evaluation[d]	C2: 3	C1: 8	B2: 17 B1: 6

Notes: [a]Self-reported on a 4-point scale: poor (1), good (2), very good (3), nativelike (4); [b]based on the average of participants' self-reported proficiency across the four skills; [c]computed for the 10 individuals who reported studying abroad in an English-speaking country for the purpose of language study; seven individuals reported studying abroad for 1 month or less, whereas three had studied abroad for more than 1 month (3, 4, and 7 months); [d]self-reported proficiency on the CEFR scale; on this scale, A1 and A2 are considered basic users, B1 and B2 independent users, and C1 and C2 proficient users.

Appendix B. Nonwords (neighboring consonant before (e.g. /s_/) or after the target (e.g., /_s/) vowel. Stimuli organized by minimal pairs (/iː - ɪ, æ - ʌ, ɜː - e, ɒ - ɔː/)

	iː - ɪ				æ - ʌ			ɜː - e			ɒ - ɔː	
Non-words (trained)	s_	siːg	sɪg	s_	sæʒ	sʌʒ	s_	sɜːg	seg	s_	sɒʒ	sɔːʒ
	z_	ziːg	zɪg	z_	zæg	zʌg	z_	zɜːg	zeg	z_	zɒg	zɔːg
	f_	fiːm	fɪm	f_	fæʒ	fʌʒ	f_	fɜːg	feg	f_	fɒʒ	fɔːʒ
	v_	viːg	vɪg	v_	væd	vʌd	v_	vɜːg	veg	v_	vɒb	vɔːb
	ʧ_	ʧiːg	ʧɪg	ʧ_	ʧædʒ	ʧʌdʒ	ʧ_	ʧɜːg	ʧeg	ʧ_	ʧɒd	ʧɔːd
	ʤ_	ʤiːd	ʤɪd	ʤ_	ʤæd	ʤʌd	ʤ_	ʤɜːd	ʤed	ʤ_	ʤɒd	ʤɔːd
	_s	jiːs	jɪs	_s	jæs	jʌs	_s	bɜːs	bes	_s	ʧɒs	ʧɔːs
	_z	jiːz	jɪz	_z	jæz	jʌz	_z	dɜːz	dez	_z	vɒz	vɔːz
	_f	jiːf	jɪf	_f	jæf	jʌf	_f	gɜːf	gef	_f	ʧɒf	ʧɔːf
	_v	jiːv	jɪv	_v	bæv	bʌv	_v	gɜːv	gev	_v	bɒv	bɔːv
	_ʧ	giːʧ	gɪʧ	_ʧ	gæʧ	gʌʧ	_ʧ	gɜːʧ	geʧ	_ʧ	mɒʧ	mɔːʧ
	_ʤ	giːʤ	gɪʤ	_ʤ	gæʤ	gʌʤ	_ʤ	gɜːʤ	geʤ	_ʤ	vɒʤ	vɔːʤ
Non-words (untrained)	s_	siːʃ	sɪʃ	s_	sæθ	sʌθ	s_	sɜːʃ	seʃ	s_	sɒf	sɔːf
		siːb	sɪb		sæʤ	sʌʤ		sɜːm	sem		sɒv	sɔːv
	z_	ziːʃ	zɪʃ	z_	zæʧ	zʌʧ	z_	zɜːθ	zeθ	z_	zɒt	zɔːt
		ziːn	zɪn		zæv	zʌv		zɜːb	zeb		zɒn	zɔːn
	f_	fiːθ	fɪθ	f_	fæʧ	fʌʧ	f_	fɜːt	fet	f_	fɒʧ	fɔːʧ
		fiːv	fɪv		fæm	fʌm		fɜːv	fev		fɒv	fɔːv
	v_	viːʃ	vɪʃ	v_	væk	vʌk	v_	vɜːp	vep	v_	vɒθ	vɔːθ
		viːð	vɪð		væm	vʌm		vɜːm	vem		vɒg	vɔːg
	ʧ_	ʧiːʧ	ʧɪʧ	ʧ_	ʧæs	ʧʌs	ʧ_	ʧɜːθ	ʧeθ	ʧ_	ʧɒt	ʧɔːt
		ʧiːd	ʧɪd		ʧæn	ʧʌn		ʧɜːb	ʧeb		ʧɒv	ʧɔːv
	ʤ_	ʤiːk	ʤɪk	ʤ_	ʤæp	ʤʌp	ʤ_	ʤɜːp	ʤɜːb	ʤ_	ʤɒp	ʤɔːp
		ʤiːv	ʤɪv		ʤæð	ʤʌð		ʤɜːg	ʤeg		ʤɒm	ʤɔːm

https://doi.org/10.1017/S0272263125101083 Published online by Cambridge University Press

Appendix C. Real words (target vowels organized by minimal pairs; spelling of target vowel in italics)

iː - ɪ				æ - ʌ				ɜː - e				ɒ - ɔː			
ee	sheep	*i*	ship	*a*	back	*u*	buck	*ir*	dirt	*e*	debt	*o*	shot	*or*	short
ea	reach	*i*	rich	*a*	cat	*u*	cut	*ur*	burst	*e*	best	*a*	what	*ar*	wart
ea	leave	*i*	live	*a*	drag	*u*	drug	*ir*	bird	*e*	bed	*o*	cod	*or*	cord
ee	seen	*i*	sin	*a*	mad	*u*	mud	*ur*	turn	*e*	ten	*a*	was	*ar*	wars

Cite this article: Nagle, C., Mompean, J. A., & Fouz-González, J. (2025). Does timing matter?: Exploring different windows of maximal opportunity to enhance the effectiveness of high variability phonetic training. *Studies in Second Language Acquisition,* 47: 1044–1070. https://doi.org/10.1017/S0272263125101083