Applied Psycholinguistics (2025), **46**, e14, 1–26
doi:10.1017/S0142716425000141

CAMBRIDGE UNIVERSITY PRESS

ORIGINAL ARTICLE

Comparing lower and higher variability multi-talker perceptual training

Charlie Nagle[1], Shelby Bruun[1] and Germán Zárate-Sández[2]

[1]Department of Spanish and Portuguese, The University of Texas at Austin, Austin, TX, USA and [2]Department of Spanish, Western Michigan University, Kalamazoo, MI, USA
Corresponding author: Charlie Nagle; Email: cnagle@austin.utexas.edu

(Received 22 August 2024; revised 17 February 2025; accepted 30 March 2025)

Abstract

One of the main challenges individuals face when learning an additional language (L2) is learning its sound system, which includes learning to perceive L2 sounds accurately. High variability phonetic training (HVPT) is one method that has proven highly effective at helping individuals develop robust L2 perceptual categories, and recent meta-analytic work suggests that multi-talker training conditions provide a small but statistically reliable benefit compared to single-talker training. However, no study has compared lower and higher variability multi-talker conditions to determine how the number of talkers affects training outcomes, even though such information can shed additional light on how talker variability affects phonetic training. In this study, we randomly assigned 458 L2 Spanish learners to a two-talker or six-talker HVPT group or to a control group that did not receive HVPT. Training focused on L2 Spanish stops. We tested performance on trained talkers and words as well as several forms of generalization. The experimental groups improved more and demonstrated greater generalization than the control group, but neither experimental group outpaced the other. The number of sessions experimental participants completed moderated learning gains.

Keywords: second language learning; speech perception; speech training; variability

Introduction

Second language learners often struggle to perceive the sounds of their additional language (L2) accurately. This difficulty is due in part to the timing of L2 learning relative to native language (L1) learning. In adult language learning, perception has been optimized for the L1, which means that adult L2 learners may equate L2 sounds with existing L1 categories despite subtle yet important crosslinguistic differences between the L1 and L2 sound systems (Best & Tyler, 2007; Flege & Bohn, 2021). This can lead to challenges in perceiving L2 contrasts accurately, which might also affect learners' ability to accurately and efficiently recognize L2 words. Some perceptual learning may occur naturally, as learners gain experience with the L2, possibly because

as their vocabulary expands, so too does their need to differentiate a larger and larger set of minimally contrastive words. Yet, the extent to which this occurs naturally, and the speed at which it occurs, depend on a range of factors, including the nature and difficulty of the learning task (Best & Tyler, 2007). In an instructed setting, targeted speech perception training can be provided to support this process by helping learners become sensitive to L2 phonetic cues and category boundaries.

Perceptual training has been shown to be effective at helping learners improve their discrimination and identification of L2 sounds (Sakai & Moorman, 2018), though the extent to which such training is associated with robust generalization and retention depends on a range of factors (Rato & Oliveira, 2023). One technique that has proven particularly effective is high variability phonetic training (HVPT), a technique which has its origin in a series of early studies training Japanese speakers on the English /l/-/ɹ/ contrast (Logan et al., 1991; Lively et al., 1993). That early work suggested an advantage for training sets in which the target contrast was produced by multiple talkers across multiple phonetic contexts. Since then, HVPT research has flourished. Several research syntheses have provided insight into the state of the art in HVPT (Barriuso & Hayes-Harb, 2018; Thomson, 2018), while addressing fundamental questions about the benefit of high variability training for speech perception (Zhang, Cheng, & Zhang, 2021) and speech production (Uchihara et al., 2024). These syntheses have also shed light on the variables that might moderate the effectiveness of this training technique, while controlling for potential confounds. Yet, pooled meta-analytic estimates are based on methodologically diverse studies. In fact, it has been relatively rare for researchers to compare several HVPT conditions within a single experimental design, though studies of that nature have begun to appear in the literature and have great potential to advance our collective understanding of how the benefits of HVPT can be further enhanced and optimized. For instance, Fouz-González and Mompean (2022) compared HVPT using symbols and keywords, targeting response options as a potential moderator of gains. In another study, Cebrian et al. (2024) examined the use of HVPT with identification and discrimination tasks, examining the training task as a potential moderator of gains. In these studies, crucially, all other elements of research methodology were held constant, which allowed for precise insight into the target variable. To contribute to the body of literature on how HVPT can be made more effective, in this study we examined how the number of talkers included in the training affects learning. Crucially, rather than comparing single and multi-talker conditions, we compared two multi-talker conditions, a two-talker, lower variability condition and a six-talker, higher variability condition.

Background

Overview of high variability perceptual training

In its most typical form, HVPT is implemented as a forced-choice identification task with trial-by-trial feedback where the learner hears a stimulus, selects a response from a closed set of options, and receives right-wrong feedback (Thomson, 2018). Discrimination tasks have also been used and have been shown to be effective (Cebrian et al, 2024), but they remain less common than identification tasks. Several modifications have been made to auditory, identification-based HVPT, which can

https://doi.org/10.1017/S0142716425000141 Published online by Cambridge University Press

be considered the baseline training format. For instance, in most HVPT research, the stimulus is a naturally produced word or nonsense word, but in some previous studies, the stimuli have been acoustically manipulated to exaggerate target features, with the goal of more effectively directing learners' attention to those features (e.g., Thomson, 2011). In another modification, auditory HVPT has also been combined with visual input, on the view that visual cues may help learners develop robust perceptual representations for targets that have a salient visual component (e.g., Hazan et al., 2005). In some cases, both acoustic exaggeration and audiovisual input (Zhang, Cheng, Qin et al., 2021) are incorporated into HVPT, and HVPT has also been combined with production training to create a more comprehensive suite of perception and production training exercises (Mora et al., 2022). As these examples illustrate, there has been a long line of productive inquiry into the ways in which HVPT might be made more effective, where effectiveness has been quantified in terms of pre-post(-delayed) gains in both perception and production (Sakai & Moorman, 2018; Uchihara et al., 2024; Zhang, Cheng, & Zhang, 2021).

Many variables could moderate the effect of perceptual training on perceptual learning, such as blocking versus interleaving talkers and contexts (e.g., Perrachione et al., 2011), response options (Fouz-González & Mompean, 2022), the training task (Carlet & Cebrian, 2022; Cebrian et al., 2024), and the type of stimuli and presentation conditions (Mora et al., 2022). A complete review of these variables is beyond the scope of this paper. Instead, in the following section, we provide a targeted review of research into the number of talkers, which is the variable we addressed in this study and arguably a central concern for all HVPT research.

Talker variability

In Logan et al. (1991) and Lively et al. (1993), the authors compared lower and higher variability perceptual training, manipulating both the number of talkers (five versus one) and the number of phonetic contexts (many versus few) included in the training. They found a modest advantage for the higher variability groups. That initial finding sparked intense interest in HVPT, especially in the extent to which multi-talker (MT) paradigms lead to gains beyond their single-talker (ST) counterparts. The primary hypothesis guiding HVPT is that MT training promotes more robust learning by helping individuals become attuned to the phonetic markers of phonological contrast while ignoring idiosyncratic patterns associated with individual talkers (and phonetic contexts). A recent meta-analysis on this issue suggests that MT training does indeed lead to a small, but statistically reliable, benefit compared to ST training (Zhang, Cheng, & Zhang, 2021). At the same time, researchers have begun to question whether the gains associated with MT paradigms are as robust as initially hypothesized. Notably, the lower limit of the confidence interval for the MT benefit in the Zhang, Cheng, and Zhang (2021) meta-analysis was 0.08 standard deviations, which means that the MT effect might be quite small. Furthermore, when Brekelmans et al. (2022) attempted to replicate Logan et al. (1991) and Lively et al. (1993), addressing several methodological shortcomings present in the original research design, they did not find reliable evidence for an MT advantage. Instead, they noted that if an MT advantage does exist, "such a benefit is likely very small" (p. 21). At the same time, they called "for

https://doi.org/10.1017/S0142716425000141 Published online by Cambridge University Press

future work to determine how and under what circumstances variability can support and boost the efficacy of phonetic training" (p. 21), echoing similar remarks by other HVPT experts (Thomson, 2018).

Talker variability can be realized in many ways, including within and between talkers. For instance, in a typical ST training format, the talker is recorded producing multiple realizations of each target item (and target items are usually selected to include a range of phonetic environments in which the target categories occur), with the goal of never exposing the learner to the exact same production twice. In other words, the learner is exposed to multiple realizations of each target item, all of which are produced by the same talker, but those realizations are phonetically distinct. Thus, ST paradigms can be characterized as presenting the listener with robust within-talker but zero between-talker variability. In MT paradigms, the same procedure is repeated to develop a stimulus set with items produced by many individuals, in which case the amount of between-talker variability present in the stimulus set is tied to the number of talkers included. From this viewpoint, ST and MT comparisons are comparisons between null (between-talker) variability and some (between-talker) variability. However, research has yet to explore whether MT conditions with fewer or more talkers (that is, MT conditions with lower or higher between-talker variability) lead to differential learning gains. If variability is a critical aspect of learning, then it stands to reason that higher variability MT conditions could promote better learning compared to lower MT conditions.

At the same time, the alternative hypothesis (that variability is not necessary or even beneficial) could also be correct. Notably, for certain learners and/or learners at certain points in their developmental trajectories, high variability may overload their processing capabilities and, as a result, have a detrimental effect on learning, compared to low variability conditions. To that point, Perrachione et al. (2011) reported that English-speaking participants with lower aptitude for pitch perception struggled to learn a four-way tonal contrast in an interleaved condition, where from one trial to the next participants were exposed to stimuli produced by different talkers. However, when the stimuli were blocked by talker, creating a condition with low trial-by-trial variability but high overall variability, the low-aptitude participants performed better. This aligns with research showing that interleaved or mixed talker conditions tend to be more challenging for L2 learners (Antoniou et al., 2015).

In summary, variability is a multidimensional concept that can be simultaneously realized through several experimental manipulations: the number of talkers, which affects the amount of overall variability, and how they are blocked, which affects trial-by-trial variability. Similar arguments can be made for phonetic contexts and even training targets (see, e.g., Shejaeya et al., 2024). Variability benefits may depend on the learning task, learner characteristics, and developmental stage. To shed further light on the role of talker variability in HVPT, and with the additional goal of contributing to current research into how this technique can be optimized, in this study we implemented variability through the number of talkers, comparing MT conditions with two and six talkers.

Target structure: Stop consonants

We targeted L1 English speakers' perception of L2 Spanish stop consonants. We selected stop consonants because they show important crosslinguistic

https://doi.org/10.1017/S0142716425000141 Published online by Cambridge University Press

differences between English and Spanish. In both languages, the primary cue to stop consonant voicing in word-/utterance-initial stops—and therefore to the distinction between voiced and voiceless stops in that phonetic context—is voice onset time (VOT), which is an acoustic-temporal variable indexing when voicing begins relative to the release of the stop consonant (see, e.g., Lisker & Abramson, 1964). VOT is not the only cue to stop voicing, but it is the most important one in both languages, which differ with respect to the location of the category boundary.

In English, voiced and voiceless stop phonemes tend to be phonetically voiceless (for an overview of stop consonant voicing in Spanish and English, see, e.g., Zampini & Greene, 2001). In voiced stops, there is a short delay between the stop release and the onset of voicing (though voiced stops are sometimes produced with prevoicing, or vocal fold vibration during closure), whereas in voiceless stops, there is a substantial delay between stop release and voicing onset. Perceptually, this means that the crossover boundary between voiced and voiceless stops occurs at approximately 30 ms of VOT. In Spanish, on the other hand, voiced stops are phonetically voiced (voiced stops are always prevoiced) and voiceless stops are phonetically voiceless, produced with the same short delay between stop release and the onset of voicing typical in English voiced stops. Perceptually, in Spanish, the crossover boundary from voiced to voiceless stops occurs at a shorter VOT value, around 0 ms. As a result, English speakers may perceive Spanish voiced and voiceless stops as instances of English voiced stops. From a crosslinguistic standpoint, stops therefore represent a good candidate structure for examining the potential benefits of lower versus higher variability MT training.

Stops are also an ideal target structure because they represent a relatively small natural class of obstruent sounds, which have been shown to respond well to perceptual training (Sakai & Moorman, 2018). We saw the potential trainability of stops as advantageous because a secondary goal of the research was to incorporate HVPT (MT) into basic university Spanish language instruction. Thus, it was important to select a structure that could be trained quickly through a few short, at-home training sessions, in line with the homework assignments students typically complete as part of their coursework.

Importantly, several studies have documented the trajectory of stop-consonant learning in instructed language learners, providing an important body of work against which the results of the present study can be compared. Briefly stated, instructed learners' perception of Spanish stops tends to improve as students move through their communicative language training (Nagle, 2018) or participate in domestic immersion programming (Casillas, 2020). Thus, even without training, learners are likely to improve, reinforcing the view that stop consonants are likely highly trainable from the earliest stages of instruction. For this reason, we also included a control group so that we could gauge the amount of learning that might occur naturally as part of general communicative language training.

The current study

Ample research demonstrates that HVPT is effective, but many open questions remain related to how MT training can be optimized. In this study, our goal was to

https://doi.org/10.1017/S0142716425000141 Published online by Cambridge University Press

target what has arguably been the most important variable in HVPT research: the number of talkers included in the training. We therefore compared lower and higher variability MT conditions to determine if learners trained on six talkers would show better learning and generalization than learners trained on two talkers. We implemented this study in a university setting, training, and testing learners using a pre-post-delayed design.

We had the following research questions and hypotheses.

1. Does phonetic training help learners improve their perception of Spanish stops?
   We hypothesized that both experimental groups would outperform the control group given the large body of research demonstrating the benefits of HVPT. At the same time, we also expected that the control participants might show some improvement.
2. Does higher variability training lead to benefits beyond lower variability training?
   We hypothesized that higher variability training would promote better outcomes for the group trained on six talkers. Though previous research has shown that MT training can overwhelm learners under certain conditions, in this study, the target structure was simple, and training stimuli were blocked by talker. For these reasons, we did not anticipate any negative effect of higher variability MT training in the present study.

Method

Participants

Participants were students enrolled in 26 sections of first-semester Spanish language coursework at two large U.S. universities, one located in the South and one in the Midwest. Students attended six lecture hours, three times a week, at the former institution and four lecture hours, either two or four times a week, at the latter. Course meetings at both institutions adopted a communicative approach, with the majority of class time centered around functional language learning activities. Course instructors were native and highly proficient non-native speakers of Spanish.

482 students were initially enrolled in the study. We excluded five participants who were not age 18 at the time of consent. Participants self-reported known vision, speech, and hearing impairments. Beyond corrective glasses, participants reported no known impairments, so we made no exclusions on this basis. Participants reported speaking 21 different L1s. English was the most common ($n = 414$), followed by Spanish ($n = 8$) and French ($n = 4$). Languages implement stop consonant voicing in different ways. To control for this fact, which could have an impact on pretest performance and pre-post-delayed gains, we excluded data from participants who indicated that they spoke an L1 other than or in addition to English or who reported speaking another language during the first five years of life (irrespective of whether they tagged that language as one of their L1s).[1] This reduced the analyzable sample to 342 participants (109 control, 114 two-talker, and 119 six-talker). We further split the data into two sets based on the number of HVPT

https://doi.org/10.1017/S0142716425000141 Published online by Cambridge University Press

Table 1. Participant characteristics

	Any sessions (n = 342)		Six sessions (n = 185)	
	M (*SD*)	Range	*M* (*SD*)	Range
Biological age	20.37 (3.44)	18–61	20.25 (3.94)	18–61
Age Spanish	9.99 (6.06)	0–52	10.18 (6.31)	0–52
Spanish context				
At home	48		22	
In school	214		119	
Both	80		44	

Note: Any Sessions refers to participants who completed at least one session. Six Sessions refers to participants who completed all six sessions. Age Spanish = self-reported age of first exposure to Spanish. Spanish Context = the primary context(s) through which the participant learned Spanish.

sessions that participants completed: the complete data set, regardless of the number of sessions completed (n = 342), and a data set including only those experimental participants who completed all six sessions (n = 185; 109 control, 41 two-talker, and 35 six-talker). This allowed us to examine learning while controlling for the number of sessions completed and to explore the effect of the number of sessions completed on gains. Participant characteristics for both data sets are given in Table 1.

Stimuli

We used a subset of stimuli from a larger stimulus set consisting of disyllabic, stop-initial words and nonsense words produced by eight native speakers of Spanish. All nonsense words obeyed Spanish phonotactic constraints (for additional details on stimulus and talker characteristics, see the supplementary online materials). We presented bilabial stop-initial words (/b-/ and /p-/) during training, reserving alveolar stop-initial words (/d-/ and /t-/) for testing, as a means of evaluating generalization of gains to a new place of articulation.

The training stimuli were produced by a total of six talkers (3F, 3M), with each talker contributing ten /p-b/ minimal pairs. The stops occurred with each of the five Spanish vowel categories (/i, e, a, o, u/) twice, yielding a total of 20 tokens per talker (2 stops × 5 vowels × 2 items). To test the effect of lower versus higher variability, we created a two-talker and a six-talker condition. We selected these talker conditions based on Zhang, Cheng, and Zhang's (2021) meta-analysis and Thomson's (2018) narrative review, where the average number of talkers in ST versus MT studies was five and the average number of talkers in HVPT research was approximately seven. Thus, we struck a balance by selecting six talkers for the higher variability condition. These talker conditions also allowed us to ensure that the six-talker condition was far more variable than the two-talker condition. We presented 120 trials per training session. To keep the number of trials consistent across the talker conditions, we used all 20 tokens per talker in the 2-talker condition, repeating the set three times. In the 6-talker condition, we used 10 tokens per talker,

https://doi.org/10.1017/S0142716425000141 Published online by Cambridge University Press

Table 2. Testing conditions and blocks

	Trained talkers	New talkers
Trained items	Block 1	Block 1
New words: Trained place	Block 2	Block 2
New words: New place	Block 3	Block 3

repeating the set twice. The same 120 items were used for all six training sessions, but the order of presentation was randomized for each participant at each session.

For testing, we created six conditions crossing trained and untrained items and talkers, including untrained items at a novel place of articulation (/t-d/). Each testing condition consisted of 20 tokens to keep the length of the test, which was administered in class, reasonable. The trained talkers were the two talkers (1F, 1M) from the two-talker condition, who were also included in the six-talker condition. The trained talkers contributed the following stimuli: for the trained words condition, ten trained words; for the untrained words condition, ten new words at the trained place of articulation (/p-b/); for the untrained place of articulation condition, ten new words at the new place of articulation (/t-d/). We repeated this procedure with two untrained talkers (1F, 1M) to create conditions involving new talkers. We structured the testing in blocks, moving from trained talkers and words in the first block (the baseline testing condition) to increasingly demanding forms of generalization. The order of blocks was held constant, but stimuli within blocks were presented in a unique random order at each testing session. The structure of testing is schematized in Table 2.

Tasks

We built the training in Gorilla Experiment Builder (www.gorilla.sc) (Anwyl-Irvine et al., 2020). The experiment consisted of seven Gorilla sessions. In Session 1, participants completed a background questionnaire, providing basic biographical information and information on their language learning experiences (e.g., native language(s), age of first Spanish exposure, years of Spanish study). Sessions 2–7 were perceptual training sessions, during which participants completed a two-alternative forced-choice identification task. In each trial, participants heard a stimulus in Spanish, saw two options presented in standard Spanish orthography, and were asked to select the option corresponding to the word they heard. Each session began with four English practice trials to familiarize participants with the structure of the task before proceeding to the 120 Spanish target trials. We decided to block the stimuli by talker based on previous research suggesting that blocking talkers, as opposed to interleaving them, is beneficial for learning (Perrachione et al., 2011), at least in the short term. The order of talkers and the order of stimuli within the talker blocks were randomized for each participant at each session. Because participants completed the training outside of class at a time and location of their choosing, we included six attention checks to ensure that participants were not simply clicking

https://doi.org/10.1017/S0142716425000141 Published online by Cambridge University Press

through the training without attending to the stimuli. Participants saw three randomly selected images, one corresponding to an English word spoken by either a female or male talker and two distractor images. Their task was to select the image corresponding to the word they heard. The attention checks appeared at random intervals throughout the training but never on two consecutive trials.

Participants received right-wrong feedback during training. When they responded correctly, they saw an indication on the following screen before moving to the next trial. When they responded incorrectly, they were required to listen to the word again and select the correct response before advancing. Each feedback screen included a randomly selected, animated image intended to increase the enjoyment of and engagement with the task. Response options did not become available until the offset of the auditory stimulus to prevent premature responses. The position of the voiceless and voiced answer choices was held constant, but we counterbalanced the position of the correct response. Participants had four seconds to respond before the trial timed out. If the trial timed out, the participant was shown a time-out screen and asked to work more quickly. Median completion time across sessions was 16.15 minutes. At the end of each training session, participants completed a short exit questionnaire, which we do not report on here.

Testing was carried out in class. The testing format mirrored the training format, except that on testing participants did not receive feedback and talker presentation was interleaved as opposed to blocked. Participants completed a two-alternative forced-choice identification task, beginning with trained items (block 1), then moving to untrained items at the trained place of articulation (block 2), and finally to untrained items at the untrained place of articulation (block 3). Each block contained items from the two trained and two untrained talkers. The three tests were identical except for the within-block item order, which was randomized for each test. We programmed the tests into students' learning management systems (e.g., Canvas), allowing participants to complete the assessment on a device of their choice (e.g., computer, tablet, phone). As with training, we held the position of the voiceless and voiced answer choices constant while counterbalancing the position of the correct response.

Procedure

We adopted a pre-post-delayed design flanking the six-session intervention. All students enrolled in the 26 sections of first-semester Spanish at the two sites participated as part of their course requirements. Students were told they would complete a series of pronunciation enrichment activities and assessments, all of which were evaluated on a complete/incomplete basis, worth 5% of their final course grade. For ease of administration, we randomized intact course sections into the two-talker, six-talker, and control conditions at each site.

During week 1 of the experiment, instructors administered the in-class pretest using a pre-recorded audio of the task instructions and test stimuli. We asked instructors to play the audio over the classroom sound system at a comfortable volume and refrain from pausing, rewinding, or replaying any portion of the recording. Test administration took less than 10 minutes. Students who were absent

https://doi.org/10.1017/S0142716425000141 Published online by Cambridge University Press

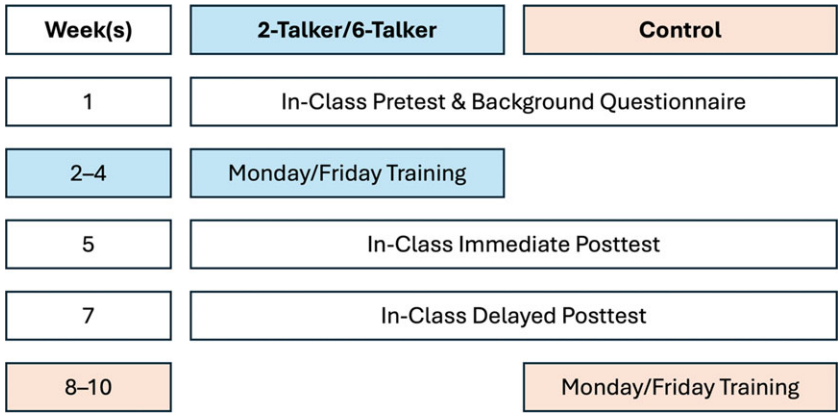

Figure 1. Experimental tasks and timeline.

on the day of testing were permitted to complete the missed assessment at home within 48 hours so that they could successfully fulfill the course requirement.

After the pretest, we provided students with a link to the Gorilla-based training system through an assignment in their learning management system. They logged into their personal profile by entering a unique code assigned to them by the research team. We used the Gorilla "delay" feature to stop students at the desired point each day and lock them out of the platform until the following session. We configured the delay such that training always took place on Mondays and Fridays to create a consistent training schedule. Participants completed the Session 1 background questionnaire the same week as the pretest. Perceptual training began on week 2 of the study, with participants training twice per week for three weeks, through week 4. We administered the in-class posttests during weeks 5 (immediate posttest) and 7 (delayed posttest) of the experiment. Control group participants completed the same training sessions but only after the conclusion of the training and testing window for the experimental groups. Figure 1 shows the experimental design and timeline.

Results

Approach to analysis

In this manuscript, we focus on the testing data. We used the *lme4* package version 1.1-35.5 (Bates et al., 2015) to fit logistic mixed-effects models to the data in R version 4.4.0 (R Core Team, 2024). First, we analyzed participants' performance on trained words and talkers. Then, we analyzed five types of generalization: (1) to new talkers, (2) to new words, (3) to new talkers and new words, (4) to a new place of articulation, which necessarily involved new words, and (5) to new talkers and new words at a new place of articulation. For the purpose of structuring the analysis, types 1 and 2 are single forms of generalization, insofar as they involve only one type of generalization, whereas types 3–5 involve generalization along multiple dimensions simultaneously. The effect structure of all models was consistent. The fixed effects of interest were

https://doi.org/10.1017/S0142716425000141 Published online by Cambridge University Press

Group (control, two-talker, six-talker), Test (pre, post, delayed), and the Group × Test interaction. We used treatment coding for both Group and Test, setting the baseline for Group to the control and the baseline for Test to pretest.

We used the *buildmer* package version 2.11 (Voeten, 2023) to create an appropriate random effects structure for data for trained words and talkers. This model then served as a benchmark for the other models we fit. We began with a sensible maximal random effects structure,[2] allowing *buildmer* to further optimize the random effects through likelihood ratio tests on nested models. We used the *DHARMa* package version 0.4.6 (Hartig, 2022) to simulate model residuals and test model assumptions; the *sjPlot* package version 2.8.16 (Lüdecke, 2024) to extract model estimates, converting log odds to odds ratios (*OR*s) for interpretability; the *emmeans* package version 1.10.2 (Lenth, 2024) to compare all groups to one another on the immediate and delayed posttests; and the *ggeffects* package version 1.7.0 (Lüdecke, 2018) to derive model-based estimates for plotting. We interpret effect sizes according to Plonsky and Oswald (2014), who recommended cutoffs of $d = .40$, .70, and 1.00 for small, medium, and large between-group effects.

Our primary analysis focused on participants who completed all six HVPT sessions, which allowed us to control for the fact that the number of sessions completed may have regulated pre-post-delayed gains. Nevertheless, we were also interested in how the number of sessions completed affected learning, leading us to undertake two additional sets of exploratory analyses to probe that issue.

Planned analyses: Six session participants

Descriptive statistics

As a first step, we examined pretest scores to see how well six-session participants performed before training and therefore how much room for learning there was. Figure 2 is a histogram of pretest means, pooling over stimulus types. As shown, scores were negatively skewed. Fifty-seven percent of participants had a mean score above 90%, suggesting that they identified Spanish stops with high levels of accuracy on the pretest, but 43% of participants had a pretest score below that threshold, suggesting that despite high overall group performance there was room for improvement for many individuals.

Next, we computed means and standard deviations by group, test, and stimulus type (Table 3). Means for the experimental groups generally increased from the pretest to the immediate posttest and then remained stable at the delayed posttest. Standard deviations for the experimental groups also decreased substantially. Interestingly, participants performed better on /t, d/ words representing an untrained place of articulation, even when combined with other forms of generalization. The new place of articulation conditions also showed lower overall variance relative to the other conditions.

Trained words and talkers (baseline)

The best model of the baseline trained words and trained talkers data included random intercepts for research sites, individual participants, talkers in the testing set, words in the testing set, and the phonological category to which the test item

https://doi.org/10.1017/S0142716425000141 Published online by Cambridge University Press

Table 3. Means and standard deviations by group, test, and stimulus type for six-session participants

	Control (n = 109)			Two-Talker (n = 41)			Six-Talker (n = 35)		
	Pre	Post	Delayed	Pre	Post	Delayed	Pre	Post	Delayed
Trained[a]	.85 (.36)	.86 (.34)	.87 (.34)	.84 (.37)	.95 (.22)	.95 (.22)	.85 (.36)	.95 (.22)	.94 (.24)
T[b]	.85 (.36)	.85 (.36)	.84 (.37)	.85 (.36)	.92 (.27)	.93 (.26)	.86 (.35)	.90 (.30)	.90 (.29)
W[c]	.91 (.29)	.89 (.31)	.91 (.28)	.94 (.25)	.97 (.17)	.97 (.18)	.93 (.25)	.97 (.18)	.97 (.16)
T/W[d]	.88 (.33)	.87 (.34)	.88 (.33)	.90 (.30)	.93 (.26)	.95 (.21)	.90 (.30)	.93 (.25)	.96 (.19)
P/W[e]	.95 (.22)	.92 (.26)	.93 (.25)	.93 (.25)	.97 (.18)	.95 (.21)	.94 (.23)	.96 (.19)	.96 (.21)
P/T/W[f]	.93 (.25)	.95 (.23)	.93 (.26)	.95 (.22)	.96 (.19)	.96 (.19)	.95 (.21)	.98 (.14)	.97 (.18)

Notes:
[a]Trained = performance on trained place of articulation, talkers, and words.
[b]T = performance on new talkers (trained words and trained place of articulation).
[c]W = performance on new words (trained talkers and trained place of articulation).
[d]T/W = performance on new talkers and new words (trained place of articulation).
[e]P/W = performance on new place of articulation and new words (trained talkers).
[f]P/T/W = performance on new talkers and new words at a new place of articulation.

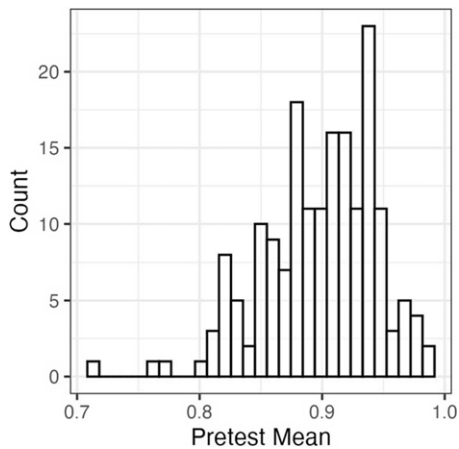

Figure 2. Histogram of pretest scores.

pertained. Simulated residuals and residual tests revealed no major issues with model fit. We therefore accepted this model as an appropriate representation of the data. As reported in Table 4, at the outset of the study, the control group already showed a strong tendency to respond correctly ($OR = 9.07, p = .003$), and the *OR*s for the simple effect of Group demonstrated that there were no statistically significant differences between the control group and the experimental groups at pretest (Control vs. Two-Talker, $OR = 0.83$, $p = .455$; Control vs. Six-Talker, $OR = 0.88, p = .691$). The *OR*s for the simple effect of test showed that the control group improved modestly over time, with the pre-delayed comparison reaching

https://doi.org/10.1017/S0142716425000141 Published online by Cambridge University Press

Table 4. Summary of final model fit to the trained talkers and words for six-session participants

	OR	*SE*	95% CI	*z*	*p*
Fixed effects					
Intercept	9.07	6.75	[2.11, 38.97]	2.96	.003
Group					
Two-Talker	0.83	0.13	[0.62, 1.12]	–1.20	.231
Six-Talker	0.88	0.14	[0.64, 1.20]	–0.82	.412
Test					
Post	1.20	0.11	[1.00, 1.44]	1.91	.056
Delayed	1.26	0.12	[1.05, 1.52]	2.42	.015
Group × Test					
2:Post	3.40	0.73	[2.24, 5.17]	5.75	< .001
6:Post	3.60	0.83	[2.29, 5.67]	5.53	< .001
2:Delayed	3.23	0.69	[2.12, 4.90]	5.50	< .001
6:Delayed	2.49	0.55	[1.62, 3.84]	4.15	< .001
Trial (covariate)	1.33	0.05	[1.24, 1.43]	7.71	< .001
Random intercepts	*SD*				
Site	0.20				
Participant	0.47				
Speaker	0.17				
Word	0.65				
Category	0.97				

Note: Model syntax: glmer(Score ~ Group*Test + scale(Trial) + (1 | Site) + (1 | Participant) + (1 | Speaker) + (1 | Word) + (1 | Category), data = data, family = "binomial," glmerControl(optimizer = "bobyqa")). Order was standardized and included as a control covariate. Group and Test were treatment-coded. For Group, baseline = Control, and for Test, baseline = Pretest.

statistical significance (posttest, $OR = 1.20$, $p = .056$; delayed posttest, $OR = 1.26$, $p = .015$). The Group × Test interaction terms were all statistically significant and > 2.00, indicating that both experimental groups improved significantly more than the control group did.

We used *emmeans* to get pairwise comparisons between groups at each time point, adjusting for multiple comparisons using the Tukey method. This analysis confirmed significant differences between the control group and the two experimental groups on the posttest with a small effect size (Two-Talker vs. Control, $OR = 2.83$, $d = 0.57$, $p < .001$; Six-Talker vs. Control, $OR = 3.16$, $d = 0.63$, $p < .001$) and on the delayed posttest with medium effect sizes (Two-Talker vs. Control, $OR = 2.69$, $d = 0.55$, $p < .001$; Six-Talker vs. Control, $OR = 2.19$, $d = 0.43$, $p < .001$). In contrast, there were no significant differences between the Six- and Two-Talker groups, and effect

https://doi.org/10.1017/S0142716425000141 Published online by Cambridge University Press

Table 5. Summary of pairwise comparisons for single forms of generalization

	Pretest			Posttest			Delayed posttest		
	OR	*d*	*p*	*OR*	*d*	*p*	*OR*	*d*	*p*
Talker									
2 vs. C	0.92	−0.05	.853	2.14	0.42	< .001	2.62	0.53	< .001
6 vs. C	1.08	0.04	.885	1.76	0.31	.006	1.90	0.35	.002
6 vs. 2	1.18	0.09	.676	0.82	−0.11	.671	0.72	−0.18	.352
Word									
2 vs. C	1.47	0.21	.190	4.09	0.78	< .001	2.89	0.59	< .001
6 vs. C	1.34	0.16	.396	3.47	0.69	< .001	3.60	0.71	< .001
6 vs. 2	0.91	−0.05	.936	0.85	−0.09	.878	1.24	0.12	.812

Note: *OR* = odds ratio; *d* = Cohen's *d* for between-group comparison. 2 = Two-Talker, 6 = Six-Talker, C = Control. Tukey-adjusted *p* values are reported to account for multiple comparisons.

sizes were small to negligible on both posttests (immediate: $OR = 1.12$, $d = 0.06$, $p = .910$; delayed: $OR = 0.82$, $d = -0.11$, $p = .704$). Figure 3 plots model-estimated probabilities for performance on trained words and trained talkers (at the trained place of articulation) in the upper left panel.

Single forms of generalization

We fit the best training data model to the generalization data sets to keep the model consistent. For the sake of space, we report on pairwise comparisons here, but full modeling details are accessible in the replication package. As reported in Table 5, there were no significant differences in performance between any of the groups on the pretest. On the posttest and delayed posttest, however, there were significant differences between the experimental groups and the control group. The *OR*s for these comparisons indicate that participants in the experimental groups were more likely than control group participants to respond correctly on the posttests. Values were consistent with a small effect size on both posttests. When the two experimental groups were compared, there was no clear advantage for either group, and the associated effect sizes were negligible. For both forms of generalization (to new talkers and new words), the mean model-estimated probability of a correct response was high overall but nevertheless showed an upward trajectory for the two experimental groups (Figure 3). The significant pairwise comparisons show that the experimental groups had a significantly higher probability of responding correctly than the control group did on both posttests.

Multiple forms of generalization

We carried out the same tests for items involving multiple types of generalization (Table 6). Results for new words and new talkers at a trained place of articulation were in line with previous results. Namely, the Two-Talker and Six-Talker groups significantly outperformed the Control group on the posttests with small effect sizes.

https://doi.org/10.1017/S0142716425000141 Published online by Cambridge University Press

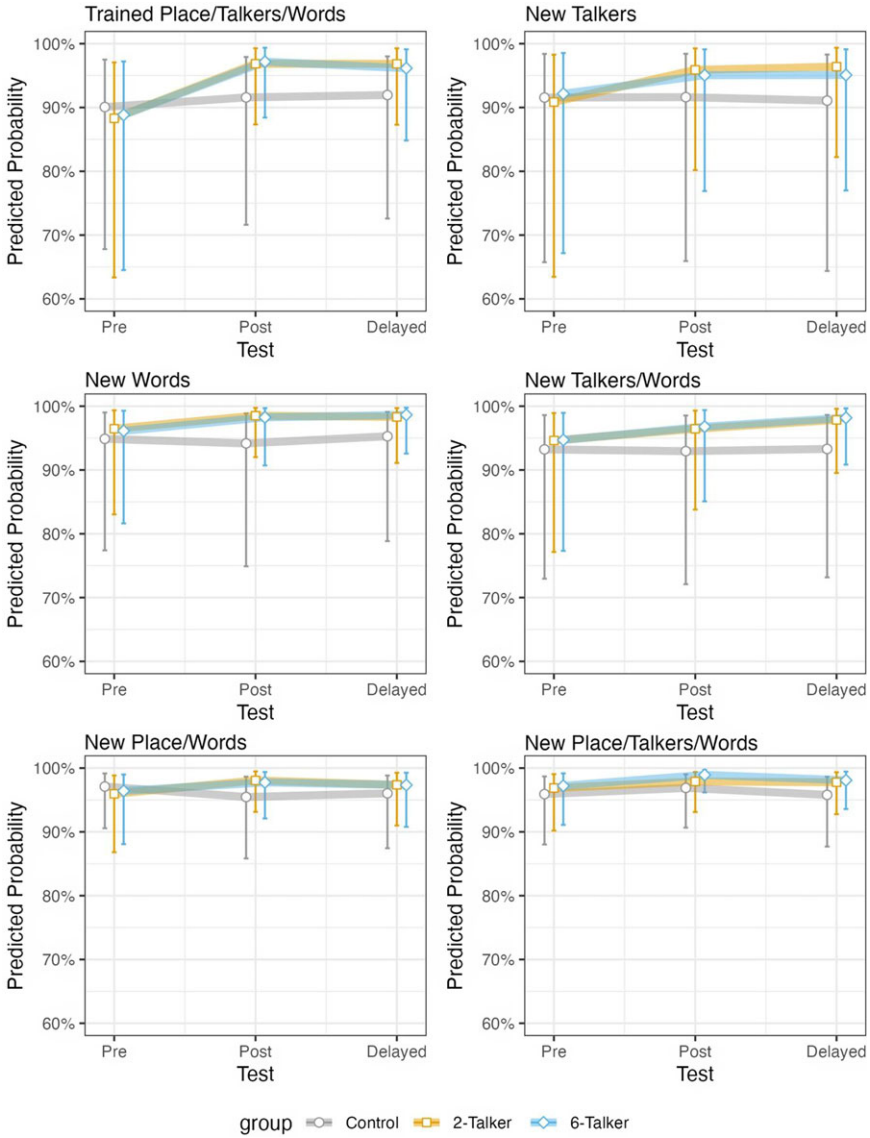

Figure 3. Predicted probabilities by group and test across conditions.

The results for (new) words at a new place of articulation showed a slightly different pattern. For trained talkers at the new place of articulation, the experimental groups significantly outperformed the Control group on the immediate posttest, with small effect sizes, but there were no statistically significant differences between the Control and the experimental groups on the delayed posttest. For new talkers at the new place of articulation, the Six-Talker group significantly outperformed the Control on both posttests, whereas the Two-Talker group outperformed the Control only on

https://doi.org/10.1017/S0142716425000141 Published online by Cambridge University Press

Table 6. Summary of pairwise comparisons for multiple forms of generalization

	Pretest			Posttest			Delayed Posttest		
	OR	*d*	*p*	*OR*	*d*	*p*	*OR*	*d*	*p*
Talker/Word									
2 vs. C	1.28	0.14	.449	2.06	0.40	.003	3.26	0.65	< .001
6 vs. C	1.30	0.14	.422	2.29	0.46	.001	3.86	0.74	< .001
6 vs. 2	1.02	0.01	.998	1.11	0.06	.922	1.18	0.09	.858
Place/Word									
2 vs. C	0.71	−0.19	.285	2.40	0.48	.003	1.53	0.23	.192
6 vs. C	0.80	−0.12	.631	2.07	0.40	.017	1.51	0.23	.250
6 vs. 2	1.14	0.07	.891	0.86	−0.08	.894	0.99	−0.01	.999
Place/Talker/Word									
2 vs. C	1.32	0.15	.512	1.49	0.22	.314	1.93	0.36	.038
6 vs. C	1.49	0.22	.289	2.98	0.60	.004	2.25	0.45	.016
6 vs. 2	1.13	0.07	.922	1.99	0.38	.179	1.17	0.09	.901

Note: *OR* = odds ratio; *d* = Cohen's *d* for between-group comparison. Tukey-adjusted *p* values are reported to account for multiple comparisons.

the delayed posttest. Notably, even when differences between the Control and experimental groups did not reach statistical significance, gains always favored the experimental groups. There were no statistically significant differences between the two experimental groups at any test point for any condition, and effect sizes were practically insignificant, save the immediate posttest comparison between the experimental groups for new words and new talkers at the new place of articulation.

Summary of findings: Planned analyses

Overall, despite high levels of initial performance, participants in the experimental groups showed more improvement in their ability to identify Spanish stop consonants than the Control participants did. This was true nearly across the board, considering performance on both the trained items and talkers and on the generalization conditions. There was no evidence that either of the experimental groups was superior to the other.

Exploratory analyses: Number of sessions completed

Because this was a classroom-based study where HVPT was implemented as homework, participants varied in terms of how many of the six target sessions they completed. To gain clear insight into the question of lower variability versus higher variability HVPT, we carried out planned analyses on data from participants who completed all six sessions. Yet, most participants did not complete all sessions, reflecting the reality of many classroom- or homework-based interventions. Therefore, we undertook two sets of exploratory analyses to examine how the

https://doi.org/10.1017/S0142716425000141 Published online by Cambridge University Press

Table 7. Control group vs. experimental group performance by test

	Pretest			Immediate posttest			Delayed posttest		
	OR	*d*	*p*	*OR*	*d*	*p*	*OR*	*d*	*p*
2T 1–3S	0.94	−0.03	.716	2.38	0.48	< .001	1.79	0.32	.010
2T 4–6S	0.98	−0.01	.834	3.60	0.71	< .001	2.40	0.48	< .001
6T 1–3S	1.02	0.01	.920	2.08	0.40	.001	1.62	0.27	.028
6T 4–6S	0.82	−0.11	.085	3.06	0.62	< .001	2.73	0.55	< .001

Note: All estimates reflect a comparison between the control group and the experimental group.

number of sessions participants completed affected their learning. First, we treated the number of sessions completed as an ordered factor with two levels: 1–3 and 4–6. We crossed this variable with Group to create a new Group variable with five levels: Control, Two-Talker 1–3 Sessions (2T 1–3S), Two-Talker 4–6 Sessions (2T 4–6S), Six-Talker 1–3 Sessions (6T 1–3S), and Six-Talker 4–6 Sessions (6T 4–6S). In this way, we were able to compare experimental groups with different completion profiles to the Control group.

Second, we examined the continuous effect of the number of sessions completed on learning gains. For this analysis, we excluded Control participants, all of whom completed zero sessions during the testing period. We therefore built a model with a three-way Group × Test × Completed interaction to explore how the number of sessions completed moderated posttest gains for the Two-Talker and Six-Talker groups. These additional, exploratory models were identical in all respects to the models from the planned analyses, other than incorporating the new target variable. Due to space limitations, we do not reproduce and describe the full set of analyses here. Instead, we focus on analysis for trained words and talkers (baseline) as an illustrative case. The full analysis can be reproduced using the data and R code provided with the replication package.

Categorical effect of sessions completed (new group variable)

For this analysis, we were interested if all experimental groups, regardless of the number of talkers and the number of sessions completed, showed improvement beyond the Control group. Thus, we set the Control group as the baseline against which the four new experimental groups were compared. All experimental groups improved more than the control group, but as shown in Table 7, participants who completed 4–6 sessions showed stronger improvement than participants who completed 1–3, as evidenced by larger *OR*s. Put another way, the number of sessions completed affected learning, whereas the number of talkers did not (aligning with the results for the participants who completed all six sessions). Another way to interpret this data is that completing more sessions is beneficial but not strictly necessary, given that even participants who completed fewer sessions (1–3) showed improvement beyond the Control group. For 1–3S participants, the effect size was small, whereas for 4–6S participants, it could be considered small to medium. Figure 4 plots this trend.

https://doi.org/10.1017/S0142716425000141 Published online by Cambridge University Press

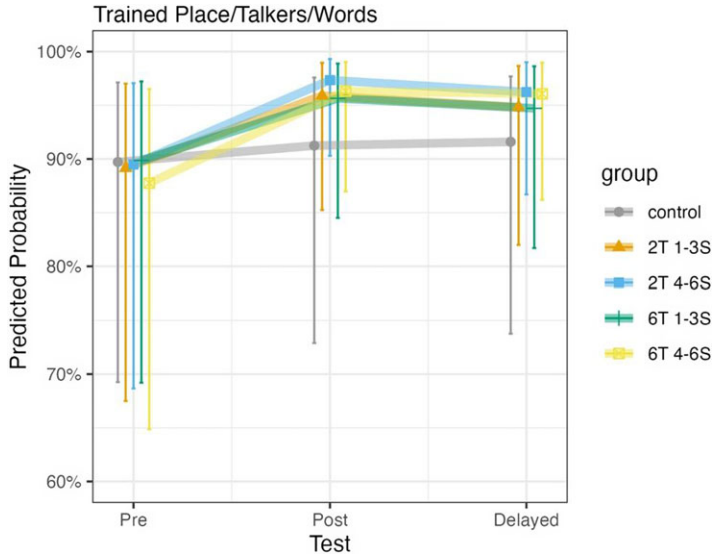

Figure 4. Performance over time by group.

Continuous effect of sessions completed (excluding control)

For this model, we set the Two-Talker group as the reference to which the Six-Talker group was compared. There was a statistically significant Test × Completed interaction. This interaction showed that the number of sessions completed had a statistically significant positive effect on immediate posttest performance ($OR = 1.18$, $SE = 0.09$, [1.01, 1.37], $p = .035$). The effect was also positive and of similar magnitude for delayed posttest performance ($OR = 1.14$, $SE = 0.08$, [0.99, 1.31], $p = .078$), but it was not statistically significant. In plain language, the more sessions that participants completed, the better performance they showed on both posttests, compared to their pretest performance. The three-way interaction with Group did not reach significance and the $OR = 1.00$, suggesting that the effect of the number of sessions completed on posttest performance was similar for the two experimental groups. Figure 5 visualizes these effects.

Summary of findings: Exploratory analyses

All experimental groups improved, regardless of the number of sessions they completed, and the number of sessions completed did not appear to have a differential impact depending on the number of talkers included in the training. Nonetheless, the number of sessions completed did affect the magnitude of improvement, such that participants who completed more sessions improved more than their peers who completed fewer sessions and showed greater improvement compared to the control group.

https://doi.org/10.1017/S0142716425000141 Published online by Cambridge University Press

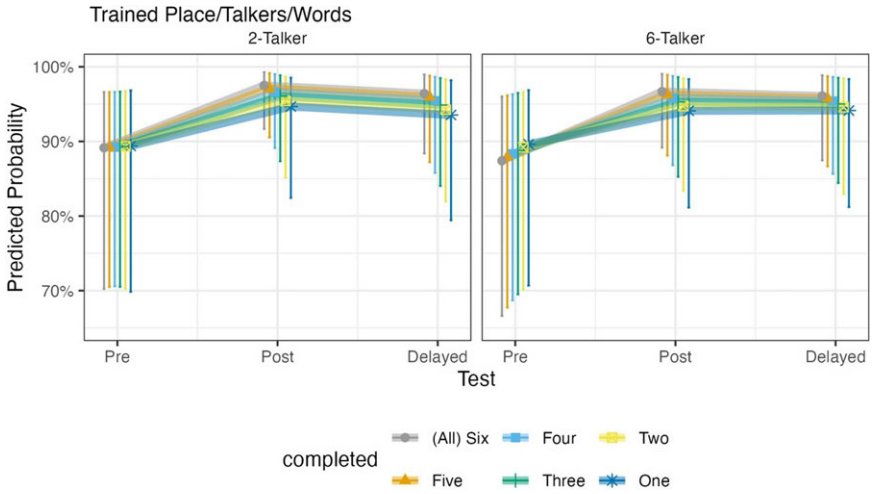

Figure 5. Effect of number of sessions completed on learning by experimental group.

Discussion

We frame the discussion in terms of the research questions that guided this study, before turning to other noteworthy findings.

RQ1: Does phonetic training help learners improve their perception of Spanish stops?

In regard to our first research question, the results of this study showed that HVPT helped learners improve their perception of Spanish stops. Although just over half of participants achieved mean pretest accuracy rates at or above 90%, the two experimental groups surpassed the Control group on the immediate and delayed posttest and showed better generalization. These findings align with previous research showing the benefits of MT phonetic training for L2 learners (e.g., Zhang, Cheng, & Zhang, 2021).

Interestingly, the benefits of HVPT were present regardless of the number of training sessions that participants completed. Even participants who completed half or less of the intended training showed evidence of learning beyond the Control group. Thus, any amount of training was beneficial for helping learners improve their identification of L2 Spanish stops. Yet, the number of sessions that participants completed did affect how much they improved, insofar as participants who completed more sessions tended to show greater gains. To the best of our knowledge, no study has explored this issue in relation to perceptual training. Nonetheless, our findings align with meta-analytic evidence showing that longer training paradigms, consisting of more sessions, are associated with greater gains than shorter paradigms (Sakai & Moorman, 2018; Zhang, Chang, & Zhang, 2021).

https://doi.org/10.1017/S0142716425000141 Published online by Cambridge University Press

At first glance, it seems surprising that there were no apparent threshold effects, where below a certain number of sessions completed, no learning was observed relative to the Control group. This may be because obstruent sounds tend to show the strongest learning gains compared to other types of sounds (Sakai & Moorman, 2018). It could also be the case that participants decided when to stop training based on their performance. In other words, perhaps participants who completed fewer sessions did so because they were already performing quite well by the second or third session and, as a result, felt that they no longer needed additional training. This could explain why any amount of training was associated with some learning. Future work could explore issues related to potential self-regulated learning.

RQ2: Does higher variability training lead to benefits beyond lower variability training?

The results of our study suggest that the Two-Talker and Six-Talker MT training conditions were equally effective at promoting learning. These results are partially in line with Brekelmans et al. (2022) in that we did not find any advantages for the higher variability condition. Crucially and differently from our study, though, Brekelmans et al. (2022) used a single talker in the low variability training, which could also be interpreted as null between-talker variability (compared to low between-talker variability for our Two-Talker group).

Undoubtedly, the difficulty of the target structure itself partially determines how much room for learning there is and the functional form that the learning curve shows. We chose to train Spanish stop consonants, which seem to be relatively easy for L2 speakers to learn and train (Sakai & Moorman, 2018). At the same time, HVPT meta-analytic research has not found the training target to be a significant moderator of training gains (Uchihara et al., 2024; Zhang, Chang, & Zhang, 2021). Nevertheless, we believe that the learning task in the present study could be considered a simple one, given that both English and Spanish show a two-category voicing distinction whose primary acoustic cue is voice onset time. Thus, for English speakers learning Spanish, learning involves recalibrating the crossover location, which according to some researchers is likely to be easier than learning tasks that involve splitting a category or creating a new one (Nagle & Baese-Berk, 2022; Vasiliev & Escudero, 2014. This may explain why participants began the study with a high level of identification accuracy and why both experimental groups performed similarly. In general, when learning involves a simple task such as boundary shift, it may be the case that any type of (MT) training, even low variability training, can promote learning. On the other hand, for more challenging learning tasks such as category split and category creation, higher variability MT conditions may be necessary or at least beneficial. In other words, the nature of the learning task, as opposed to the type of L2 target considered in isolation, may interact with training characteristics, shaping the training gains observed. Similar arguments can be made for the number of sessions completed, which in the current study did not interact with the number of talkers.

It is important to observe that training was blocked by talker for both groups, and blocking has been shown to have a significant positive impact on learning (Zhang,

https://doi.org/10.1017/S0142716425000141 Published online by Cambridge University Press

Cheng, & Zhang, 2021) and may even level the playing field when training learners with diverse cognitive profiles (Perrachione et al., 2011). This is precisely why we chose to block by talker in the present study. It is thus possible that in interleaved conditions, there would be a clearer difference between the Two-Talker and Six-Talker groups. In addition, interleaved practice may lead to more robust long-term learning (Schorn & Knowlton, 2021; Taylor & Rohrer, 2010), a form of testing that was beyond the scope of this study but is certainly worth pursuing in future research.

Generalization

We trained participants on bilabial stops (/b, p/) and tested them on bilabial and dental (/t, d/) stops. Findings for the generalization tests at the trained (bilabial) place of articulation mirrored findings for trained talkers and words: the experimental groups significantly outperformed the control group on both posttests, but neither experimental group showed better performance than the other. Findings were different for the generalization tests at the untrained (dental) place of articulation. We reasoned that dental stops would be challenging for two reasons: first, because participants were not trained on that place of articulation and second because coronal stops (an umbrella term referring to stops articulated with the front part of the tongue) have an alveolar place of articulation in English but a dental place of articulation in Spanish. Contrary to our expectations, participants performed exceptionally well on dental stops at all testing points, which could be due to the order in which test items were presented. The last block of items tested were the items related to generalization to a new place of articulation. It could be argued, then, that participants performed well on those items because they had already practiced on the preceding items corresponding to trained talkers and words (presented in the first block) and single forms of generalization (presented in the second block of items). In other words, perhaps the structure of the testing itself served as a form of training, even though the testing did not include feedback and words within blocks were presented in a random order on each test. It seems unlikely that this explanation could fully explain the results because when we analyzed the effect of Block on performance, we found a significant improvement between the first and second block, which we interpret as possible evidence of a task familiarity effect (whereby performance improved modestly as participants grew accustomed to the testing task), but not between the second and third block.[3]

Instead, high performance on /t, d/ may be related to the different place of articulation that coronal stops have in the two languages. Perhaps this difference made it easier for participants to create a distinct and robust mental category for Spanish dental stops relative to bilabial and possibly velar stops, which share the same place of articulation in both languages. To our knowledge, no study has examined differences in stop consonant perception at multiple places of articulation in the same L2 sample, but English and Spanish coronal stops have been shown to differ with respect to several acoustic parameters (Casillas et al., 2015). If learners are sensitive to these differences, then they may be more successful at creating or updating perceptual representations for coronal stops compared to stops at other places of articulation. Testing velar stops, which share the same place of articulation

https://doi.org/10.1017/S0142716425000141 Published online by Cambridge University Press

in both languages, in addition to bilabial and dental stops, could shed further light on this issue.

The (random) effect of site, section, and participant

When we evaluated the random effects portion of the model, model comparisons suggested that random intercepts for sites and participants improved fit, but intercepts for sections did not. This suggests that there was some variation in overall performance by site and by participant, while sections performed remarkably similarly. This makes sense given that the sections at each site followed a similar communicative language curriculum. Thus, any differences may have been captured by the site-level intercept as opposed to intercepts for individual sections. The estimated fixed intercept was 9.07, whereas the standard deviation for the by-site intercept was 0.20, which is relatively small compared to the fixed effect. This, in turn, suggests that there was little variation across the two sites, despite the fact that they were located in different geographic regions of the US with different sociodemographic and sociolinguistic characteristics. One site was in a region with a large population of Spanish speakers, in which case it is possible that learners at that site were at least passively exposed to Spanish outside the classroom, whereas the other site was in a region with a much smaller population of Spanish speakers, making outside exposure unlikely. Yet, these differences did not seem to drive any practically significant differences in performance and learning across sites. There was some variation in performance across participants, but the standard deviation for the participant intercepts (0.47), though larger than the standard deviation for sites, can also be considered small relative to the fixed effect.

When we attempted to model random slopes for Group, Test, and the interaction term, model comparisons suggested that none of the random slopes significantly improved model fit. In plain terms, this means that trajectories were highly similar across the grouping units present in the clustered data. Put another way, there was limited variation between participants in terms of learning. Relative homogeneity of trajectories could be related to the way we implemented HVPT, insofar as our training methods may have encouraged a similar amount of learning for all participants. Another possible explanation is the target structure. Obstruents are known to respond well to training (Sakai & Moorman, 2018), and L2 Spanish stops do not seem to be a particularly difficult L2 contrast for L2 English speakers, whose L1 shares the same number of phonological categories and relies on the same phonetic cue to implement them. Thus, we expect that with other, more challenging training targets, by-participant slopes for Test (Time) could reveal substantial heterogeneity with respect to the amount of improvement observed over time.

Limitations and future directions

This study has several limitations. First, even though we implemented the training in the first semester of intensive university-level Spanish language coursework, many participants were already near ceiling. We found that the experimental groups' performance improved, but ceiling effects may have made it difficult to observe a difference between the Two-Talker and Six-Talker conditions.

https://doi.org/10.1017/S0142716425000141 Published online by Cambridge University Press

Furthermore, we trained and tested participants under optimal listening conditions: stimuli were words presented in isolation in the absence of noise. Thus, although the present results shed light on generalization, they cannot speak to how robust and resilient participants' perceptual categories were. Testing participants in noise could mitigate ceiling effects while also providing insight into how accurately participants perceive the target contrast under more realistic listening conditions. Embedding the contrast into sentences[4] or evaluating reaction time could further illuminate training-induced changes in the L2 perceptual system.

Despite participants' high overall initial performance, we hope that this study can serve as a blueprint for assessing the specific impact of potential HVPT moderators in a single experimental design. Meta-analyses can shed light on aggregate effects, considering the entire body of evidence, which can in turn guide future experimentation, but the studies included in any meta-analysis differ on many design features, making it difficult to gain precise insight into specific variables. In this study, we compared lower and higher variability training considering the number of talkers, but features such as blocking and interleaving talkers, the number of sessions, and even how talkers are presented and blocked or interleaved across sessions (e.g., blocking during the first few sessions before moving to an interleaved format), could be targeted in future work. Future research could also adopt the current experimental design with L1-L2 pairs that imply different learning tasks, such as category split and category formation, to gain additional insight into how the number of talkers affects learning under those scenarios. We also hope to have provided a model for how future work can systematically approach classroom-based research, where, for instance, participants may not fully comply with intended training protocols and timelines. Ultimately, we believe that the true litmus test for HVPT and other training paradigms is how well such paradigms work despite deviations like skipping a session or two. In that regard, examining training trajectories and the factors that influence them is another important avenue for future research, and we are in the process of doing so for the data collected in this study.

Conclusion

In this study, we specifically targeted the number of talkers as a potential moderator of HVPT gains. We compared Two-Talker and Six-Talker conditions, asking whether lower and higher variability MT would lead to differential learning patterns, training English speakers on L2 Spanish stops. We found that both experimental groups improved compared to the Control group. In exploratory, post hoc analyses, we also observed that the experimental groups improved regardless of how many sessions they completed, but the number of sessions appeared to regulate the amount of improvement observed. Put another way, the number of sessions did not seem to affect if improvement was observed, but rather how much was observed. This suggests that even minimal training can be beneficial, at least for relatively simple learning scenarios such as the one investigated in the current study. Future work should continue to explore the impact of specific variables on the efficacy of HVPT and the extent to which the optimal format for HVPT depends on the nature of the learning task, potentially connecting such work with the notion of desirable difficulty in L2 practice (e.g., Suzuki et al., 2019).

https://doi.org/10.1017/S0142716425000141 Published online by Cambridge University Press

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/S0142716425000141.

Replication package. All research materials, data, and analysis code are available at https://osf.io/mnks9/.

Acknowledgments. We would like to thank the participants of the Fall 2023 Undergraduate Research Assistantship Cohort on the Science of Language Learning, whose help was instrumental in getting this project up and running. Thank you to the language program directors and coordinators for their help and guidance in implementing the project and to the instructors who welcomed us into their language courses. We are also grateful to Ron Thomson for providing constructive feedback on an early version of this manuscript and to the reviewers for their thoughtful comments.

Author 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, project administration, writing – original draft, writing – review and editing. Shelby Bruun: conceptualization, data curation, investigation, methodology, project administration, writing – original draft, writing – review and editing. Germán Zárate-Sández: data curation, investigation, project administration, writing – original draft, writing – review and editing.

Funding. This material is based upon work supported by the National Science Foundation under Grant Nos. 2309561 and 2414107.

Competing interests. The authors declare none.

Notes

1 We did not exclude participants below a certain threshold for their self-reported age of first exposure to Spanish, nor did we exclude participants who indicated that their primary context for learning Spanish was at home because upon reviewing information about years of exposure to Spanish, it became clear that the quantity, quality, and consistency of exposure to Spanish was highly variable. Thus, we felt that any exclusion strategy in relation to age of first exposure would be arbitrary and imprecise. To that point, neither variable was significantly related to pretest performance after we filtered participants who reported an L1 other than or in addition to English or who reported speaking another language during the first five years of their life.

2 The model that we fed into *buildmer* was: buildmer(score ~ group*test + scale(item_order) + (1 + group*test | site) + (1 | section) + (1 + test | participant) + (1 | word) + (1 + group*test | speaker) + (1 + group*test | category). We refer to this model as a sensible maximal model because it contains a sensible set of random slopes, based on our hypotheses about where variation might be observed in the data. Specifically, we fitted by-site, by-participant, by-speaker, and by-category random slopes for test because we thought that sites, participants, speakers, and phonetic categories might show varying gains over time. We fit by-site, by-speaker, and by-category random slopes for the group*test interaction because we thought that the extent to which each group improved over time could vary across sites, speakers, and phonetic categories. For instance, it could be the case that one of the groups (such as the 6-talker group) improved more on one of the speakers who produced the stimuli, which would be captured by the by-speaker random slopes for the interaction term. By the same token, we did not fit by-section random slopes for the interaction because we thought that sections at each site would show similar trajectories based on their shared curriculum and teaching methodology. Thus, we reasoned that variation would be best captured by the by-site random effects. Likewise, we had no strong reason to believe that certain words would improve more than others or that certain groups would show better performance or learning on certain words. We therefore did not fit any by-word random slopes.

3 A simple model with Block as a predictor showed that participants performed significantly better on block 2 compared to block 1, but other comparisons did not reach statistical significance: block 1 vs. 2, $OR = .64$, $p = .039$; block 1 vs. 3, $OR = .51$, $p = .696$; block 2 vs. 3, $OR = .78$, $p = .955$.

4 In Spanish, in many contexts, phonologically voiced stops weaken to approximants, which means that if sentences were used as stimuli, stops would need to occur in both word-initial and sentence-initial positions. In other contexts, additional cues to stop consonant voicing contrasts beyond the ones relevant to the stimuli in this study may come into play.

References

Antoniou, M., Wong, P. C., & Wang, S. (2015). The effect of intensified language exposure on accommodating talker variability. *Journal of Speech, Language, and Hearing Research*, **58**(3), 722–727. https://doi.org/10.1044/2015_JSLHR-S-14-0259

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavioral Research Methods*, **52**, 388–407. https://doi.org/10.3758/s13428-019-01237-x

Barriuso, T. A., & Hayes-Harb, R. (2018). High variability phonetic training as a bridge from research to practice. *The CATESOL Journal*, **30**(1), 177–194.

Bates, D., Maechler, M., Boker, S., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In M. J. Munro & O.-S. Bohn (Eds.), *Second language speech learning: The role of language experience in speech perception and production* (pp. 13–24). 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**(2), 271–299. https://doi.org/10.1017/S0142716421000515

Casillas, J. V. (2020). The longitudinal development of fine-phonetic detail: Stop production in a domestic immersion program. *Language Learning*, **70**(3), 768–806. https://doi.org/10.1111/lang.12392

Casillas, J. V., Díaz, Y., & Simonet, M. (2015). Acoustics of Spanish and English coronal stops. In The Scottish Consortium for ICPHs 2015 (Ed.), Proceedings of the 18th International Congress of Phonetic Sciences. The University of Glasgow.

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

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**(2), 297–328. https://doi.org/10.1017/s0272263120000455

Hartig, F. (2022). DHARMa: Residual diagnostics of hierarchical (multi-level/mixed) regression models (Version 0.4.6) [R package]. https://CRAN.R-project.org/package = DHARMa

Hazan, V., Sennema, A., Iba, M., & Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication*, **47**(3), 360–378. https://doi.org/10.1016/j.specom.2005.04.007

Lenth, R. (2024). emmeans: Estimated marginal means, aka least-squares means (Version 1.10.4) [R package]. https://CRAN.R-project.org/package = emmeans

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, **20**(3), 384–422. https://doi.org/10.1080/00437956.1964.11659830

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. *The Journal of the Acoustical Society of America*, **94**(3), 1242–1255. https://doi.org/10.1121/1.408106

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/and/l/: A first report. *The Journal of the Acoustical Society of America*, **89**(2), 874–886. https://doi.org/10.1121/1.1894649

Lüdecke, D. (2018). ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, **3**(26), 772. https://doi.org/10.21105/joss.00772

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

Mora, J. C., 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), 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**(1), 234–270. https://doi.org/10.1111/lang.12275

Nagle, C. L., & Baese-Berk, M. M. (2022). Advancing the state of the art in L2 speech perception-production research: Revisiting theoretical assumptions and methodological practices. *Studies in Second Language Acquisition*, **44**(2), 580–605. https://doi.org/10.1017/s0272263121000371

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**(1), 461–472. https://doi.org/10.1121/1.3593366

Plonsky, L., & Oswald, F. L. (2014). How big Is "big"? Interpreting effect sizes in L2 research. *Language Learning*, **64**(4), 878–912. https://doi.org/10.1111/lang.12079

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

Rato, A., & Oliveira, D. (2023). Assessing the robustness of L2 perceptual training: A closer look at generalization and retention of learning. In U. Kickhöfel Alves & J. I. Alcantara de Albuquerque (Eds.), *Second language pronunciation: Different approaches to teaching and training* (pp. 369–398). De Gruyter Mouton.

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**(1), 187–224. https://doi.org/10.1017/S0142716417000418

Schorn, J. M., & Knowlton, B. J. (2021). Interleaved practice benefits implicit sequence learning and transfer. *Memory & Cognition*, **49**(7), 1436–1452. https://doi.org/10.3758/s13421-021-01168-z

Shejaeya, G. A., Roon, K. D., & Whalen, D. H. (2024). Talker variability versus variability of vowel context in training naïve learners on an unfamiliar class of foreign language contrasts. *Journal of Phonetics*, **107**, 101369. https://doi.org/10.1016/j.wocn.2024.101369

Suzuki, Y., Nakata, T., & Dekeyser, R. (2019). The desirable difficulty framework as a theoretical foundation for optimizing and researching second language practice. *The Modern Language Journal*. https://doi.org/10.1111/modl.12585

Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. Applied Cognitive Psychology, 24(6), 837–848. https://doi.org/10.1002/acp.1598

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). *Journal of Second Language Pronunciation*, **4**(2), 208–231. https://doi.org/10.1075/jslp.17038.tho

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**(4), 591–623. https://doi.org/10.1017/S0142716424000195

Vasiliev, P., & Escudero, P. (2014). Speech perception in second language Spanish. In K. L. Geeslin (Ed.), *The Handbook of Spanish Second Language Acquisition* (pp. 130–145). Wiley.

Voeten, C. (2023). buildmer: Stepwise elimination and term reordering for mixed-effects regression (Version 2.11) [R package]. https://CRAN.R-project.org/package = buildmer

Zampini, M. L., & Green, K. P. (2001). The voicing contrast in English and Spanish: The relationship between perception and production. In J. Nicol (Ed.), *One mind, two languages*. Blackwell.

Zhang, X., Cheng, B., Qin, D., & Zhang, Y. (2021). Is talker variability a critical component of effective phonetic training for nonnative speech? *Journal of Phonetics*, **87**, 101071. https://doi.org/10.1016/j.wocn.2021.101071

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**(12), 4802–4825. https://doi.org/10.1044/2021_JSLHR-21-00181

Cite this article: Nagle, C., Bruun, S., & Zárate-Sández, G. (2025). Comparing lower and higher variability multi-talker perceptual training. *Applied Psycholinguistics*. https://doi.org/10.1017/S0142716425000141