Studies in Second Language Acquisition (2025), 1–34
doi:10.1017/S0272263125100879

RESEARCH ARTICLE

High variability phonetic training (HVPT): A meta-analysis of L2 perceptual training studies

Takumi Uchihara[1], Michael Karas[2] and Ron I. Thomson[3]

[1]Graduate School of International Cultural Studies, Tohoku University, Sendai, Japan; [2]Department of Applied Linguistics, Brock University, St. Catharines, ON, Canada and [3]Department of Applied Linguistics, Brock University, St. Catharines, ON, Canada
Corresponding author: Takumi Uchihara; Email: takumi@tohoku.ac.jp

(Received 22 October 2022; Revised 07 April 2025; Accepted 21 April 2025)

Abstract

This meta-analysis of 79 studies evaluates the effectiveness of high variability phonetic training (HVPT) for the development of second language (L2) speech perception and explores learner-related and methodological variables that influence training effects. The overall medium-to-large effects of HVPT on L2 speech perception support the effectiveness of HVPT, for both pretest-posttest comparison (g = 0.92, k = 96) and treatment-control comparison (g = 0.67, k = 32), confirm long-term retention of perception gains, and, to some extent, indicate generalization of learning to novel stimuli. Training effects are influenced by several key variables (length of L2 learning, response labels, type of training task, type of testing task, total training time, target phones, and number of talkers). The findings provide compelling evidence to support the efficacy of HVPT for L2 perceptual learning and suggest circumstances under which training effects are optimized.

Keywords: high variability phonetic training (HVPT); L2 perceptual learning; L2 pronunciation; L2 speech perception; meta-analysis

Introduction

In second language (L2) speech acquisition, developing the ability to perceive L2 sounds accurately is paramount because it underlies recognition of spoken words (Melnik & Peperkamp, 2021) and leads to increased comprehension of longer stretches of spoken discourse (Vandergrift & Baker, 2015). The acquisition of accurate L2 speech perception also anchors the development of L2 speech production skills (Sakai & Moorman, 2018; Uchihara, Karas, & Thomson, 2024). Except in highly artificial learning environments (e.g., Sheldon & Strange, 1982), accurate speech production is contingent on the ability to reliably perceive contrasts between L2 phonemes (see Thomson, 2022 for an overview). Among numerous approaches to improving L2 perception skills

(e.g., shadowing, imitation), high variability phonetic training (HVPT) is increasingly considered the most empirically supported phonetic training paradigm in L2 speech research (Thomson, 2018; Uchihara et al., 2024). HVPT entails perceptual training with trial-by-trial feedback in which learners are trained using multiple instantiations of the target sounds (e.g., segmentals) produced by multiple talkers (voices) in varied phonetic contexts. HVPT studies have shown perception improvement for L2 liquids (Logan, Lively, & Pisoni, 1991), vowels (Lambacher, Martens, Kakehi, Marasinghe, & Molholt, 2005; Thomson, 2012b), stops (Flege, 1995b), fricatives (Lengeris & Nicolaidis, 2014), tones (Wang, Spence, Jongman, & Sereno, 1999), and syllable structure (Huensch & Tremblay, 2015). Benefits of HVPT have generalized to the perception of trained sounds in untrained phonetic contexts (Carlet & Cebrian, 2022) and produced by untrained talkers (Herd, Jongman, & Sereno, 2013). Retention of gains has lasted from two weeks (A. H. Lee & Lyster, 2016) to six months (Silpachai, 2020) and has led to improvement in the production accuracy of target sounds (Bradlow et al., 1997, 1999; Thomson, 2011; Uchihara et al., 2024).

A longstanding issue in HVPT research is that its practical value has not been widely translated to use by language teachers and learners. Thomson (2018) attributes this lack of uptake to the fact that the majority of HVPT studies are published in technical phonetics journals, which are mostly inaccessible to those who would most benefit from this knowledge. This is unfortunate given robust evidence that learners who complete a course of HVPT training almost always experience immediate and statistically significant gains in their perception of target sounds (e.g., usually 10 to 15%) targeting various phonological features and different target languages (Flege, 1995b; Iverson, Pinet, & Evans, 2012). Despite the demonstrable usefulness of this training technique, HVPT does not necessarily bring the same level of improvement in every study. For example, perception gain scores have ranged from 5% (Qian, 2018) to 29% (Yeon, 2004). Numerous studies have demonstrated that HVPT-induced perception improvement is modulated by a range of variables, including learner characteristics (e.g., perception abilities, Perrachione, Lee, Ha, & Wong, 2011); L2 experience (Iverson et al., 2012), training formats (e.g., blocking talker by individual training sessions vs. intermixing talkers within each training session, Fuhrmeister & Myers, 2020); and the nature of training tasks (e.g., identification vs. discrimination, Carlet & Cebrian, 2022). To promote the integration of HVPT in L2 instructional contexts, researchers need to provide clear guidance concerning how it can be optimized across a variety of learning contexts (e.g., for whom and how the training should be implemented).

Given a massive increase in the number of HVPT studies, especially over the last decade or so, the time is ripe to conduct a large-scale meta-analysis in this area. The primary goal of the current meta-analysis is twofold. First, we aim to comprehensively establish HVPT's efficacy, by examining the true benefit of HVPT for developing L2 speech perception (i.e., *Does HVPT lead to substantial perception gain?*), retention (i.e., *Are gains retained for an extended period after training?*), and generalization (i.e., *Is the gain robust enough to accurately perceive L2 sounds produced by an untrained talker in untrained phonetic contexts?*). Second and more central to knowledge translation, we examine under what circumstances this training technique brings about the greatest improvement by examining the impacts of learner-related and methodological variables. We hope that this will help expand the use of HVPT and its associated speech learning principles to a much broader range of learning contexts than is currently the case.

HVPT and L2 perceptual learning

Canonical HVPT has been defined as perceptual training comprising three necessary features: 1) variability in talkers, 2) variability in the phonetic contexts (or words) in which target sounds are presented, and 3) the provision of corrective feedback (Thomson, 2018). Many other training features vary across HVPT studies (e.g., inclusion of production training, see Mora Ortega, Mora-Plaza, & Aliaga-García, 2022), but these three features are essential for training to meet the definition of HVPT. A central tenet of HVPT is that it simulates an L2 immersion environment, where learners receive varied input of L2 sounds, with the added benefit of explicit feedback on the accuracy of perception, which is mostly not available in the real world. HVPT is compatible with an exemplar view of speech perception in which talker- and item-specific information are encoded and stored. Such a simulated input-rich condition maximizes the potential for input to be integrated, such that listeners develop stable and robust phonetic categories (Pierrehumbert, 2002; Zhang, Cheng, & Zhang, 2021). In turn, this facilitates accurate perception and production of the L2 sounds, regardless of the age at which speakers start learning the L2 (Flege, 1995a; Flege & Bohn, 2021).

Logan et al.'s (1991) seminal study was the first to demonstrate the effectiveness of HVPT for improving the perception accuracy of English liquids (/r/ vs. /l/) by first language (L1) Japanese listeners. In this study, six L1 Japanese learners listened to 68 English minimal pairs that contained either an /l/ or an /r/ in various positions, produced by five talkers. During this training, learners completed a two-alternative forced-choice identification task (i.e., pressing a button to indicate whether they heard a minimal pair containing either an /l/ or /r/), with trial-by-trial feedback on their performance. Learning was assessed through an identification task before and after 15 training sessions in a pretest-posttest design with the same test items. The post-training test was followed by a generalization task, which tested their perception of the same target sounds using items and talkers that did not occur in training. Results showed significant improvement from pretest (78.1%) to posttest (85.9%). Furthermore, improvement in trained items at posttest generalized to untrained stimuli produced by trained talkers. However, listeners were less accurate in their identification of /l/ and /r/ when novel stimuli were produced by a novel talker (i.e., a talker not heard during training or testing). In a follow-up study, Lively, Pisoni, Yamada, Tohkura, and Yamada (1994) also confirmed that gains in /l/-/r/ perception were fully maintained after a three-month delay. While performance had regressed by the six-month interval, it remained significantly better than the pretraining level.

With a slightly different focus, Lively, Logan, and Pisoni (1993) compared the effect of exposure to multiple talkers with a single talker to investigate the unique contribution of talker variability as a single source of variability. The authors conclude that participants' capacity to generalize to untrained stimuli produced by a novel talker was attributed to their exposure to training stimuli produced by multiple talkers rather than a single speaker (but see Brekelmans, Lavan, Saito, Clayards, & Wonnacott, 2022 replication for contradictory findings). A recent meta-analysis conducted by Zhang et al. (2021) shows that there is a small but significant advantage for the multiple-talker condition over the single-talker condition with respect to generalization to novel talkers.

Subsequent studies have revealed that HVPT enhances the perceptual learning of consonants beyond the /l/-/r/ contrast as well as vowels. Lambacher et al. (2005) confirmed that the treatment group's identification of five English vowels (produced by multiple talkers) improved more than the control group. In that study, Japanese

university students completed a five-alternative forced-choice identification training task for six weeks, while the control group received no training. The treatment group showed significant improvement in perception from pretest to posttest, whereas the control group did not. Subsequent studies have demonstrated similar effects for treatment over control groups for a wide variety of target sounds including Arabic glottal and pharyngeal fricatives (Burnham, 2013–2014), Portuguese bilabial stops (de Oliveira, 2020), English vowels (Aliaga-Garcia, 2017), and Spanish intervocalic consonants (Herd et al., 2013), among many others.

Our review of the relevant literature indicates that HVPT consistently improves perception accuracy (e.g., pretest vs. posttest performance; treatment vs. control conditions). In fact, perception improvement in HVPT studies has almost universally been significant and is attributed to variation in talker and phonetic context. In contrast, it should be noted that Brekelmans et al.'s (2022) study provides contradictory evidence that the use of multiple talkers does not promote superior gains relative to exposure to a single talker. Brekelmans et al.'s study certainly raises questions that need to be further investigated. They conclude that variability is essential to learning, but that variability in phonetic contexts is more important than variability in talkers. As noted earlier, perception gains vary considerably across studies, indicating that the effect of talker variability and contextual variability is moderated by myriad other variables. In what follows, we provide an overview of variables known to impact the effectiveness of HVPT.

Factors that influence the effectiveness of HVPT in perception improvement

Research has revealed numerous factors influencing the development of L2 speech perception. For example, individual differences have been shown to play some role. L2 proficiency and experience have some impact, such that advanced learners with rich L2 experience in L2 immersion contexts may benefit less from HVPT (Iverson et al., 2012; Wong, 2014). Such learners may have naturally developed more accurate category representations because of a richer array of phonetic experience. Thus, a few hours of computer-based training may not help them further enhance their category knowledge as much as it improves the knowledge for less experienced learners, whose performance is likely to be far below the ceiling. The precise extent to which L2 experience and proficiency influence learning outcomes of HVPT remains to be determined, with previous studies showing inconsistent findings: e.g., no difference in perceptual learning between experienced and inexperienced learners (Iverson et al., 2012; Wong, 2014), a larger training effect for learners with lower proficiency (H.-Y. Lee & Hwang, 2016), and a larger effect for learners with L2 immersion experience (Georgiou, 2021). Age differences may also play a role in training effectiveness. Although Flege (1995a) and Flege and Bohn (2021) propose that the general perceptual mechanisms for learning L2 phonetic categories remain intact over a life span, differences in the degree of improvement across ages have been observed (Giannakopoulou, Brown, Clayards, & Wonnacott, 2013; Shinohara & Iverson, 2021).

Another crucial consideration is the target sounds. Much of the early work on HVPT involved training Japanese adults to distinguish English /l/ and /r/ (e.g., Bradlow et al., 1997; Lively et al., 1993; Logan et al., 1991). This was followed by studies involving learners from different L1s and focused on other consonantal contrasts (e.g., Catalan speakers learning English stops, Carlet, 2017) as well as vowels (e.g., Lambacher et al., 2005; Nishi & Kewley-Port, 2007; Thomson, 2012b). With this expansion of targets, researchers have also investigated the potential benefits of HVPT for spontaneous

improvement in production. Examining how general perceptual training, including HVPT, enhances L2 production learning, Sakai and Moorman's (2018) meta-analysis demonstrated that training effects were larger for obstruents than for vowels or sonorants. These differences may come from different levels of processing required to perceive obstruents (categorical perception) vs. vowels (continuous perception). The latter is said to be more difficult (Carlet & Cebrian, 2022). Yet, this claim is contradicted by Zhang et al.'s (2021) and Uchihara et al.'s (2024) meta-analyses of HVPT studies, which did not find a significant difference between learning vowels versus consonants. Besides segmentals, HVPT can also improve the perception of suprasegmentals, such as Chinese lexical tone learning (e.g., Silpachai, 2020; Y. Wang et al., 1999) and English syllable structure (e.g., Korean adults learning English palatal codas: CVC vs. CVCy with "y" representing the palatal, Huensch & Tremblay, 2015).

The choice of training and testing tasks is another critical consideration in HVPT research. Researchers normally train and test learners with an identification task, a discrimination task, or a combination of both. In an identification task, learners hear an auditory stimulus and select the sound/word that they think they heard from a set of labelled responses (e.g., *read*, *lead*). In a category discrimination task, learners are aurally presented with sets of stimuli produced by different talkers and required to indicate whether two sounds are same or different (i.e., AX discrimination), whether a second sound is similar to the first or a third sound (i.e., AXB discrimination), whether the third sound is similar to the first or second sound (i.e., ABX discrimination), or which of the three sounds is different from the remaining two sounds (i.e., oddity discrimination). Identification tasks are believed to involve higher levels of phonological encoding and top-down processing of speech signals in which listeners respond based on their phonetic representations in memory. Discrimination tasks tend to utilize a lower-level sensory mode of perception that detects relevant acoustic cues for phonetic distinction (Carlet & Cebrian, 2022; de Oliveira, 2020; Iverson et al., 2012; Leong, Price, Pitchford, & Heuven, 2018). Although some researchers have argued that identification training is a more appropriate task for improving the perception accuracy of L2 sounds (Jamieson & Morosan, 1986; Logan & Pruitt, 1995; Strange & Dittmann, 1984), evidence supporting the efficacy of discrimination training has also been increasing (e.g., Carlet & Cebrian, 2022; Cebrian, Gavaldà, Gorba, & Carlet 2024; Flege, 1995b; Shinohara & Iverson, 2018). Comparison studies examining whether there is a difference between the two are needed if the goal is to optimize training.

Many other modifications to canonical HVPT have been attempted, leading to further insights and speculation concerning the best approach to use. These have included examining the optimal number of target sounds (full vowel sets vs. a subset comprising only difficult vowels, Nishi & Kewley-Port, 2007), types of corrective feedback (provision of wrong signal only, target sounds, nontarget sounds, or target +nontarget sounds, A. H. Lee & Lyster, 2016), stimulus type (nonwords vs. real words, Thomson & Derwing, 2016), response labels (keywords vs. phonetic symbols, Fouz-González & Mompean, 2021), trial-by-trial talker presentation (blocked vs. intermixed, Fuhrmeister & Myers, 2020), and adaptive training (adaptive vs. fixed HVPT, Yang, Nanjo, & Dantsuji, 2021). Furthermore, several studies have explored how the provision of additional information about training stimuli affects the effectiveness of HVPT. For example, researchers examined the impact of adding a visual modality to perceptual training, so that learners can see the talkers' faces (Hardison, 2003), and/or their hand gestures (Hirata & Kelly, 2010). Synthetic manipulation of auditory stimuli has also been used to provide learners with more salient cues to target sounds (e.g., doubling length of vowels, Thomson, 2011), which might orient learners' attention to the key

acoustic properties that signal a difference between target segments (e.g., Iverson, Hazan, & Bannister, 2005; X. Wang & Munro, 2004).

The present study

This study adopts a meta-analytic approach to synthesize a large number of previous HVPT studies aimed at improving L2 perception accuracy. Prior to this current work, there are three related meta-analyses (Sakai & Moorman, 2018; Uchihara et al., 2024; Zhang et al., 2021), but with slightly different foci. Sakai and Moorman (2018) focused on perceptual training more generally (i.e., not limited to HVPT) and examined improvement in production. Similarly, while Uchihara et al. (2024) conducted a meta-analysis of 31 studies focusing exclusively on HVPT, they only examined production gains, not those for perception, for which there are many more studies. Uchihara et al. (2024) report significant and small-to-medium effects of HVPT on gains in L2 production accuracy (g = .49 and .66), although they did not find strong support for long-term retention nor generalization to untrained stimuli. The current meta-analysis focuses exclusively on the effect of HVPT on L2 speech perception. While Zhang et al. (2021) also focused on perceptual learning, the goal of our meta-analysis is different from theirs. Zhang and colleagues examined how a specific component of HVPT (i.e., talker variability) contributes to perceptual learning based on 18 studies, finding a significant advantage for multiple-talker input over single-talker input on immediate perception gains (g = .28, k = 13), generalization to untrained talkers (g = .36, k = 9), and retention (g = 1.09, k = 2).

In contrast, the current meta-analysis responds to Thomson's (2018) call to optimize HVPT as a standalone protocol for perceptual training in instructional contexts. It does this by examining the effectiveness of 79 canonical HVPT studies (i.e., those incorporating talker variability, context variability, and feedback), with attention to moderating design variables that differ across studies. We focus on perceptual learning outcomes to answer three questions. First, we aim to measure the overall effectiveness of HVPT for perceptual learning. HVPT effectiveness is evaluated in terms of (a) pretest-posttest immediate improvement, (b) long-term retention, and (c) generalization to untrained stimuli. Notably, the analytical procedure adopted for perception generalization is unique in that we examine both complete and partial generalization. Partial generalization refers to learning generalized to "new" test stimuli (i.e., items [words or syllables] *or* talkers that did not appear during training but did appear during pretests). Complete generalization refers to learning generalized to "novel" stimuli (i.e., items [words or syllables] or talkers that did not appear during either training or pretests). This distinction is important, given the possible impact of exposure to target stimuli during testing on the validity of generalization (de Oliveira, 2020; Logan & Pruitt, 1995). Second, we attempt to clarify the influence of individual differences and methodological choices in moderating HVPT results. This meta-analysis will not only seek to confirm the effectiveness of HVPT for perceptual learning but, more importantly, will determine under what circumstances HVPT brings about the greatest benefit for L2 learners.

Third, we examine the degree to which perception accuracy of control groups improves over time. Control group participants complete pretest and posttests either absent training (e.g., Hirata, Whitehurst, Cullings, 2007) or with training on nontarget sounds (e.g., Huensch & Tremblay, 2015). Control groups provide a baseline against which to compare HVPT treatment groups, while accounting for the influence of extramural exposure to L2 input or the practice effect from taking pretests (Carlet & Cebrian, 2022). Compared to the first two goals, the third goal does not appear directly

relevant to HVPT efficacy. However, knowing the overall magnitude of perception gains made by control groups is important to determine a baseline for expected incidental learning (e.g., extramural input or testing effects) and to interpret true HVPT effects.

The present study was guided by the following research questions:

1. How effective is HVPT for improving L2 perception accuracy?
   1a. What immediate improvement in L2 learners' perception accuracy is realized through HVPT?
   1b. Is improvement in L2 perception accuracy retained over time?
   1c. Does improvement in L2 perception accuracy generalize to novel stimuli?
2. To what extent do learner-related and methodological variables moderate the effectiveness of HVPT for improving L2 perception accuracy?
3. To what extent does L2 perception accuracy increase in untrained learner controls?

Method

Literature search

When searching for literature and screening data, this study followed the PRISMA guideline for reporting systematic reviews and meta-analyses (Page et al., 2021; Zhang, Cheng, & Zhang, 2022). Numerous comprehensive literature searches were conducted to ensure all relevant manuscripts were included in the meta-analysis. This study included journal articles, book chapters, conference proceedings, master's theses and doctoral dissertations, published since 1991, the year of Logan et al.'s (1991) seminal study. The first search was conducted in early 2021, and two qualitative syntheses on HVPT (Barriuso & Hayes-Harb, 2018; Thomson, 2018) were also consulted for relevant studies. Keywords were drawn from these initial articles to search the following relevant databases: Linguistics and Language Behavior Abstracts (LLBA), ProQuest Dissertations, ProQuest Education, Education Resources Information Center (ERIC), Web of Science, and Psycinfo. The first search looked for matches in manuscript titles, abstracts and/or keywords and included the following keyword combinations: ("HVPT" or "high variability phonetic training" or "high variability perceptual training" or "computer assisted pronunciation training" or "high variability segmental perceptual training") and ("L2 phonetic training" or "L2 speech perceptual training") and ("identification task" or "discrimination task"). After querying databases, searches of 11 disciplinary journals were also conducted using the same search terms but this time allowing keywords to appear in the title, abstract, keywords and/or full manuscript: *Journal of Phonetics*, *Journal of the Acoustical Society of America*, *Language Learning*, *Phonetica*, *Language Learning & Technology*, *Computer Assisted Language Learning*, *ReCALL*, *Annual Review of Applied Linguistics*, *Bilingualism: Language and Cognition*, *Studies in Second Language Acquisition*, and *Applied Psycholinguistics*. These journals were selected following Sakai and Moorman (2018) and based on the search results from Thomson's (2018) review. Still, as a further layer, table of contents from five key journals (*Applied Psycholinguistics, Journal of Phonetics*, *Journal of the Acoustical Society of America, Phonetica,* and *Phonology*) were manually searched between 1991 and 2021 to ensure no manuscripts were missed. Finally, searches were conducted on Google and Google Scholar, again allowing search terms to appear in either the title, abstract, keywords and/or manuscript, and pages were reviewed until saturation. As the review and revision process took time, database searches were conducted again in July of 2023 using the same keywords and databases noted above, but this time restricting

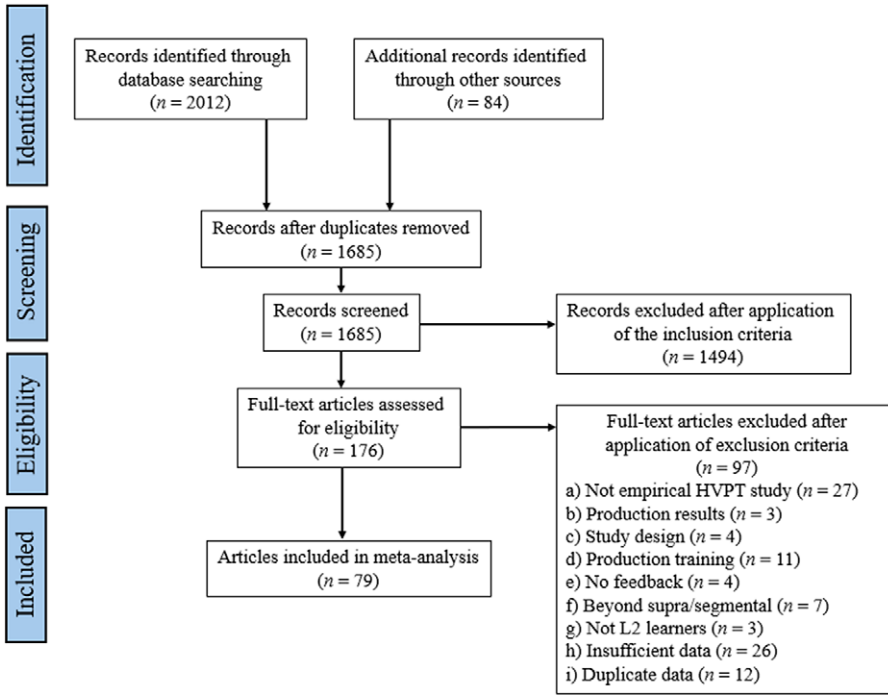

Figure 1. Literature scan for HVPT studies.

the manuscripts to be from 2020 onwards. See Figure 1 for literature search process summary.

Inclusion and exclusion criteria

We identified a total of 2,096 studies as potentially eligible to be included in the current meta-analysis. Two coders (the first and second authors) screened the initial manuscripts and removed duplicates leaving 1,685 manuscripts. A further 1,494 manuscripts were removed after applying the following inclusion criteria:

(a) The study focused on perceptual HVPT defined as perceptual training with three key features held intact (i.e., talker variability, phonetic context variability, and trial-by-trial feedback), as outlined by Thomson (2018). Studies that included only a single talker who provided training stimuli were not included. Furthermore, studies that did not provide immediate feedback were not included.
(b) The study operationalized HVPT as required in criterion (a); however, the canonical version of HVPT could also be extended upon and include techniques intended to enhance perceptual learning, such as audiovisual input, adaptive training, and/or acoustic cue manipulation.
(c) The study was an empirical investigation of perceptual HVPT contrasting either pretest versus posttest perception performance (i.e., within-participant design) or treatment versus control conditions (i.e., between-participant design).

(d) The target of HVPT was L2 speech sounds including segmental sounds (i.e., vowels and consonants) and prosodic features (e.g., tone) measured through speech perception tasks such as identification and discrimination tasks.
(e) The study targeted L2 learners, without any reported learning disability, above primary school level. Participants could have any L1 background.
(f) The report was written in English.

After applying the inclusion criteria, 176 manuscripts were retained and the full texts were reviewed, resulting in a further 97 reports being excluded for one of the following reasons.

(a) Studies that were not empirical HVPT investigations were omitted. HVPT was defined as perceptual training in which learners received auditory stimuli from two or more talkers in various phonetic environments with immediate feedback (Thomson, 2018). Based on this, 27 studies were removed because they only included a single talker for perceptual stimuli or did not adhere to canonical HVPT training (e.g., including distractor memory tasks during HVPT, Antoniou & Wong, 2015).
(b) Because this meta-analysis focused on perception improvement, 3 studies that included only production results were excluded (e.g., Wiener, Chan, & Ito, 2020).
(c) Four studies were excluded because of study design. For example, in Iino (2019), specific pre- and posttests were not used. Instead, initial and final training sessions were used as tests in place of pre- and posttests; thus, Iino (2019) and similar studies were removed from analysis.
(d) We focused exclusively on perceptual training. Thus, if studies included production training as part of HVPT, they were removed from analysis. This resulted in 11 studies being excluded.
(e) A key element of HVPT is trial-by-trial feedback for learners. Thus, four studies were removed because they did not include immediate feedback.
(f) A further seven studies were excluded because they investigated linguistic elements beyond segmental or suprasegmental features (e.g., vocabulary, Barcroft & Sommers, 2005; grammar, Bulgarelli & Weiss, 2021).
(g) Manuscripts were removed if they did not focus on L2 learners or included learners with impairments. This resulted in 3 studies being removed.
(h) Studies were removed if they did not include sufficient data for statistical information (e.g., sample size, means, standard deviations, t value). This resulted in 26 studies being removed.
(i) Finally, studies were removed if they contained data that was in other sources. This occurred occasionally as some published work drew on graduate theses. A further 12 reports were excluded to avoid duplication of data.

Coding

From the included studies, data were extracted and coded for the following information: (a) study characteristics (e.g., authors, year, publication type), (b) learner profiles (participants' L1, target language, country, learning context [foreign language vs. second language], proficiency [novice vs. beginner vs. lower-intermediate vs. intermediate vs. upper-intermediate vs. advanced], age of learning, and age of testing), (c) stimuli features (target phone [vowel vs. obstruent vs. sonorant vs. syllable structure vs. tone], inclusion of nonwords [yes vs. no]), (d) training features (task type [identification vs. discrimination],

response label [keyword vs. orthography vs. phonetic symbol vs. visual image], corrective feedback type [target vs. target and nontarget combined vs. wrong signal only], talker presentation [blocked vs. intermixed], environment [laboratory vs. participant controlled vs. classroom], adaptive training [adaptive vs. fixed], number of talkers, number of target phones, number of phonetic contexts, and number of response choices), (e) testing features (test type [identification vs. discrimination] and test item and talker type [old item-old talker vs. old item-new talker vs. new item-old talker vs. new item-new talker]), (f) modifications to HVPT (audiovisual input [yes vs. no] and acoustic cue manipulation [yes vs. no]), (g) training intensity (duration [> two months vs. one to two months vs. one week to one month vs. < one week], number of sessions, number of trials per session, total number of trials, training time per session, and total training time), and (h) results of interest (e.g., mean and standard deviation of the pre- and posttest for treatment and control groups). For detailed information about each coding category, see Description of Coding Categories in Appendix 1 in the Supplementary Materials.

Initially, the first two authors independently coded the studies. To ensure accuracy of coding for key variables (i.e., proficiency, number of target sounds, corrective feedback type, talker presentation, training environment, and test item and talker type), a second round of coding was completed. The interrater reliability between the two coders was 100% for number of sounds and talker presentation, 99.64% for test item and talker type, 98.02% for proficiency, 97.09% for training environment, and 95.33% for corrective feedback type. When agreement could not be reached on certain items, the third author served as the tiebreaker. The completed coding sheet is available via https://osf.io/qc8hr/.

Effect Size Calculation

We adopted Hedges' g as the basic unit of analysis, the transformed version of Cohen's d, which corrects for bias in small samples (Borenstein, Hedges, Higgins, & Rothstein, 2009). Regarding the calculation of effect size for within-participant data, studies reporting sample means and standard deviations for pretest and posttest performance were extracted from a pool of descriptive data. The gain score was calculated by subtracting the pretest score from posttest score, and the standard deviation for the gain score was computed using Equation 1 (see Appendix 2 for Equations 1 to 11 used to calculate effect sizes in this meta-analysis). Because the pretest-posttest correlation was needed to compute the standard deviation for the gain score, we calculated the mean correlation based on eleven studies providing raw scores ($N_{participant}$ = 164) and imputed an estimated pretest-posttest correlation of .60 to compute the standard deviation and sampling variance. In calculating the standardized difference in pretest and posttest means (Cohen's d) and its sampling variance (Equation 3), as suggested by Boreinstein et al. (2009), we followed Equation 2 to adjust the standard deviations for the gain scores using the pretest-posttest correlation. For studies not reporting sample means and standard deviations for pretest and posttest performance, t values (Equation 4) or sample mean differences and standard deviations for the differences (Equation 5), were used to calculate effect sizes. The resultant effect sizes were converted to Hedges' g with its sampling variance (see Equation 6).

Because some studies reported multiple descriptive data from the same participants (i.e., from both identification and discrimination tests, performance on multiple target phones, and outcomes based on different item and talker types), multiple scores for individual learners were averaged to yield a composite score. This avoided violating the

requirement of independence of observations. To consider the nested structure of multiple data, we averaged the observed effect sizes and aggregated the sampling variances for the effect sizes using the *agg* function in the *MAd* package (Del Re & Hoyt, 2010). Since this method requires the correlation between the multiple effect sizes, we first followed Zhang et al. (2021) to impute the correlation at r = .50 as the starting point. We then confirmed the robustness of the results irrespective of different imputed values, i.e., r = .25 and r = .75 (see Appendix 3 for the results of sensitivity analyses). The same procedure was applied for the calculation of other effect sizes from within-participant designs (i.e., pretest vs. delayed posttest, posttest vs. generalization posttest, and control group's pretest vs. posttest).

For calculation of the effect size for the between-participant data, we extracted studies reporting the sample means and standard deviations for both treatment and control groups. The group mean differences were calculated by subtracting the control-gain scores from the treatment-gain scores. Following Morris (2008), we used pretest data and computed the pooled pretest standard deviation (Equation 7) to calculate Cohen's d with its sampling variance (Equation 8), later converted to Hedges' g (Equation 9). When two or more means and standard deviations were available from multiple independent treatment groups, the sample size weighted mean and the pooled standard deviation across groups (Equation 8) were calculated. For studies not reporting summary statistics, F values were used to compute effect size estimates (Equation 10).

Statistical analysis

We employed the Comprehensive Meta-Analysis software (Borenstein et al., 2022) and used a random-effects model to compute the weighted mean effect size, assess between-study heterogeneity, and conduct moderator analyses. The original data for analysis are accessible via https://osf.io/qc8hr/. The analysis for RQ1 was conducted to determine the overall effectiveness of HVPT for perceptual learning in terms of immediate improvement (RQ1a), retention (RQ1b), and generalization (RQ1c). Regarding RQ1a (immediate improvement), we conducted two separate analyses according to different study designs to produce the within-participant weighted effect size (i.e., the mean difference between pretest and posttest performance or the gain score) and the between-participant weighted effect size (i.e., the mean difference of gain scores between HVPT treatment and control groups). A total of 99 effect sizes were available for the analysis of the within-participant effect size and 35 for the analysis of the between-participant effect size.

Regarding RQ1b (long-term retention), we computed the weighted standardized mean difference (a) between posttraining accuracy scores and delayed posttest scores and (b) between pretest scores and delayed posttest scores. Studies reporting delayed posttest scores were selected for this analysis (k = 30). The mean interval between posttests and delayed posttests was 2.3 months (SD = 1.8, range = 5 to 6 months). Regarding RQ1c (generalization to novel stimuli), we computed the weighted standardized mean difference between posttest scores and generalization posttest scores. Studies reporting generalization posttest scores (i.e., test items and/or talkers that did not appear during training or pretests) were selected for this analysis. The resulting breakdown of available studies included categories of novel item and novel talker (k = 12), novel item and old talker (k = 14), and old item and novel talker (k = 5).

To answer RQ2 regarding the influence of moderator variables on the effectiveness of HVPT, we used a mixed-effects model to conduct moderator analysis with 15 categorical and seven continuous variables. Subgroup analyses were conducted with a

between-group Q statistic for predetermined categorical variables. For continuous variables, meta-regression analyses were conducted with a full Maximum Likelihood method. The moderator analyses were conducted only for the within-participant data given the statistical robustness with the larger sample size (k = 99). Listwise deletion was applied to deal with missing data (for the number of effect-size samples available for the moderator analysis, see Appendix 1).

To answer RQ3 regarding the improvement of control groups, we computed the weighted mean difference between the pretest and posttest accuracy scores for control groups. Studies reporting pretest and posttest scores for control groups taking tests and receiving perceptual training on nontarget sounds, and those taking tests without receiving any training were selected for this analysis (k = 31).

We evaluated the impact of between-study heterogeneity using the Cochrane Q statistic (i.e., test of significance) and the I^2 statistic (i.e., proportion of the observed variance reflecting variance in true effects rather than sampling error). While the Q statistic is sensitive to sample size, I^2 statistic is not (Boreinstein et al., 2009), and its value on the order of 25%, 50%, and 75% can be interpreted as low, moderate, and high degrees of heterogeneity (Higgins, Thompson, Deeks, & Altman, 2003). Given the limitation of the I^2 statistic (i.e., not indicating the variability of the true effects on an absolute scale), we additionally reported the prediction interval, which indicates how much the effect size varies in the target population (Boreinstein, 2022). The magnitude of the effect size was interpreted according to Plonsky and Oswald's (2014) L2 field-specific benchmarks for independent standardized mean differences (.40 for small, .70 for medium, and 1.00 for large effects). To examine the influence of potential outliers and ensure the stability of the results, we conducted a leave-one-out analysis, which performs the calculation of the mean effect size multiple times by excluding one study at each analysis. Thus, we can investigate the influence of each study on the overall effect-size estimate and identify influential studies (see Appendix 4 for the results of leave-one-out analyses). Publication bias was evaluated through the construction of a funnel plot and analyzed using Egger's test. In instances where Egger's test suggested the presence of publication bias, the trim-and-fill method was employed to adjust for asymmetry.

Results

Description of included studies

The 79 included studies comprised journal articles (n = 54), doctoral dissertations (n = 14), conference proceedings (n = 8), a master's thesis (n = 1), and a book chapter (n = 1). Out of 54 journal articles, a relatively small number of articles (n = 13) were published in journals focusing on applied linguistics or language teaching (e.g., *Studies in Second Language Acquisition*, *Language Learning*, *Language Learning & Technology*). The majority of the articles (n = 41) were published in technically oriented journals focusing mainly on speech learning and psycholinguistics, such as *The Journal of the Acoustical Society of America* (n = 15), *Journal of Phonetics* (n = 5), *Journal of Speech, Language, and Hearing Research* (n = 4), and *Applied Psycholinguistics* (n = 4).

Among the included studies, 79 reports produced 99 unique experimental groups (available for pretest-posttest comparison) and 35 experimental groups that were compared with control groups (available for treatment-control comparison). The mean sample size of the 99 unique experimental groups was 17.2 (SD = 9.5), ranging from eight (Y. Wang et al., 1999) to 80 (Carlet, 2017). A small correlation between sample size

and publication year (r = .24) indicates no clear increase in sample size by year. No studies except Qian (2018) reported the test reliability (e.g., Cronbach's alpha) for either pretest or posttest perception measures. The vast majority of experiments focused on English as a target language (k = 66), followed by Mandarin Chinese (k = 12), Japanese (k = 12), French (k = 2), Korean (k = 2), and other languages (Arabic, Greek, Hindi, Portuguese, and Spanish). Both monolingual and bilingual speakers participated in training experiments with their L1s including English (k = 25), Japanese (k = 15), Korean (k = 10), Mandarin Chinese (k = 8), Catalan-Spanish (k = 8), Thai (k = 7), Spanish (k = 5), Greek (k = 5), Cantonese (k = 4), Mandarin-Cantonese (k = 3), and other languages (e.g., Arabic, Russian, Portuguese, Malay, French). Based on the studies reporting age of learning (AOL) (k = 36) and age of testing (AOT) (k = 91), participants started learning a target language at the age of 9.7 (SD = 4.9, range = 1.5 to 19.8) and they were on average 21.4 years old (SD = 5.1, range = 7.9 to 38.8) when they participated in the experiments. A large number of experiments were conducted in foreign language (FL) contexts (k = 80) compared to second language (SL) contexts (k = 16) or both (k = 3).

The current data set of 99 unique experiments demonstrated a variety of training programs employing different training formats and procedures. In the training programs, the average number of target sounds was 6.1 (range = two to 35), and the average number of phonetic contexts was 26.6 (range = 2 to 132). The majority of studies presented stimuli produced by four talkers (k = 45), followed by two (k = 13), five (k = 12), and six (k = 10), with the average number of talkers being 5.2 (range = 2 to 30). Intensity of the training programs varied considerably across experiments: the average number of sessions was 8.9 (range = one to 45 sessions), the average number of training trials and training time per session were 256 trials (range = 48 to 720 trials) and 36.3 minutes (range = five to 120 minutes) respectively, and the average number of total training trials and total training time were 2,043 trials (range = 256 to 12,240 trials) and 294 minutes (range = 60 to 1,125 minutes) respectively. Eighty-three experiments used an identification task, seven used a discrimination task, and eight used a combination of different perception tasks.

The 79 manuscripts included 163 experimental groups providing at least information about the mean pretest-posttest difference of perception accuracy in percentage score for each of the four categories of the test-item and test-talker types. The unweighted mean difference in Table 1 descriptively shows that learners tended to achieve a higher gain on the posttest of old-item-and-old-talker type (M = 14.12%) compared to other types (the means ranging from 11.89% to 12.96%). However, because the confidence intervals for all item-and-talker types overlapped with each other, there appears to be no substantial differences between the four categories. This indicates that HVPT may work for improving the ability to perceive L2 sounds accurately, even when either test items or talkers never appeared during training. Based on the experiments reporting pretest and posttest accuracy scores (k = 155), the correlation between the two test scores was .66 ($p < .001$), and the correlation between the pretest and gain score was

Table 1. Descriptive statistics for unweighted mean differences between pretest and posttest perception accuracy (K = 163)

	k	M(%)	SD(%)	95% CI
Old item–Old talker	37	14.12	5.80	[12.81, 16.05]
Old item–New talker	36	12.41	5.85	[10.43, 14.39]
New item–Old talker	18	11.89	5.86	[8.97, 14.80]
New item–New talker	72	12.96	8.50	[10.96, 14.95]

Note: k = number of studies. CI = confidence interval.

–.52 ($p < .001$). These results indicate that HVPT is effective for learners with a range of perception abilities, but the degree to which their perception accuracy improved became smaller as learners' pretraining perception abilities were higher.

RQ1a: What immediate improvement in L2 learners' perception accuracy is realized through HVPT?

We calculated the summary effect size for the 99 samples denoting the pretest-posttest perception gains. The summary effect size was large and significant, $k = 99$, $g = 1.02$, 95% CI [.90, 1.13], $p < .001$. The heterogeneity between the studies was significant and high, $Q(98) = 522.03$, $p < .001$, $I^2 = 81.2\%$, indicating that only about 20% of the variability was due to sampling error. After we conducted a leave-one-out analysis and excluded three influential cases (see Appendix 4 for the results), the summary effect size remained large and significant, $k = 96$, $g = .92$, 95% CI [.83, 1.00], $p < .001$. The heterogeneity between the studies was slightly smaller than the one reported in the original data, yet the variability was still considered high, $Q(95) = 282.80$, $p < .001$, $I^2 = 66.41\%$. The 95% prediction interval ranged from .25 to 1.59, indicating that with 95% probability, a future observation of the training effect will fall in this interval (see Appendix 5 for the forest plot of 96 studies).

Based on the 96 effect sizes for the mean pretest-posttest difference, publication bias was first visually assessed with a funnel plot (Figure 2), indicating a tendency that studies with a lower precision (or higher SE) produced larger effect sizes. To further quantify this potential publication bias asymmetry, we conducted an Egger's test, confirming the presence of publication bias, $t(94) = 8.36$, $p < .001$. The trim-and-fill method identified 28 samples that could be hypothetically added to retrieve the symmetry shape of the data distribution, resulting in a smaller estimate, $g = .71$, 95% CI [.62, .81]. These results indicate the possibility that the observed mean effect was overestimated due to the influence of publication bias. We also performed sensitivity analyses to examine whether the summary effect size and variance estimates changed if different values for within-study correlations had been used. The results confirmed that

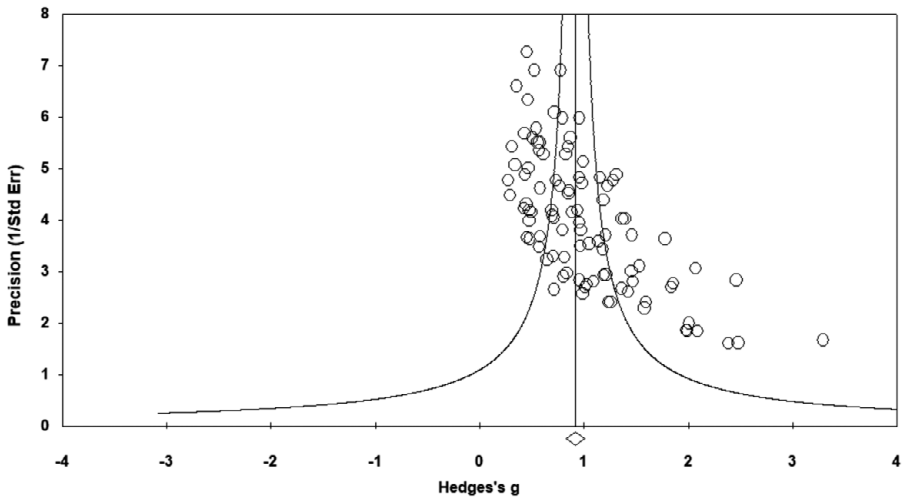

Figure 2. Funnel plot of perception effect size (mean pretest-posttest difference) by inverse standard error.

no substantial changes were observed regardless of the different values used for calculating the summary effect size (see Appendix 3 for the results of sensitivity analyses).

In addition to estimating the mean effect size for the pretest-posttest difference, we aggregated the 35 effect sizes for the difference in perception gain scores between the treatment and control groups. The summary effect size was large and significant, k = 35, g = .90, 95% CI [.67, 1.14], p < .001. The heterogeneity between the studies was significant and high, Q(34) = 152.86, p < .001, I^2 = 77.8%, indicating that only about 22% of the variability was due to sampling error. After we excluded three influential cases based on the leave-one-out analyses (see Appendix 4 for the results), the summary effect size remained significant but smaller with a medium effect, k = 92, g = .67, 95% CI [.55, .79], p < .001. The heterogeneity became much lower and nonsignificant, Q(32) = 34.35, p = .310, I^2 = 9.8%. The 95% prediction interval ranged from .43 to .92, indicating that with 95% probability, a future observation of the training effect will fall in this interval (see Appendix 5 for the forest plot of 32 studies). Publication bias was visually assessed with a funnel plot (Figure 3). No clear association between precision and effect sizes was observed, which was confirmed with the nonsignificant result of the Egger's test, t(30) = 1.64, p = .112. Sensitivity analyses confirmed that the use of different within-study correlations made no substantial changes to the summary effect size and variance estimates (see Appendix 3 for the results of sensitivity analyses).

RQ1b: Is improvement in L2 perception accuracy retained over time?

Based on 30 studies reporting perception accuracy scores at pretest, posttest, and delayed posttest, two analyses of effect-size aggregation were conducted for the comparison of pretest versus delayed posttest scores and posttest versus delayed posttest scores (see Appendix 6 for detailed results and the forest plot). The mean effect size for the pretest-delayed-posttest contrast was significant and large, g = .98, SE = .10, 95% CI [.79, 1.17], p < .001, but note that the effect size adjusted for potential publication bias was medium, g = .73, 95% CI[.53, .94] (see Appendix 6 for the analysis of publication bias). The mean effect size for the posttest-delayed-posttest contrast approached

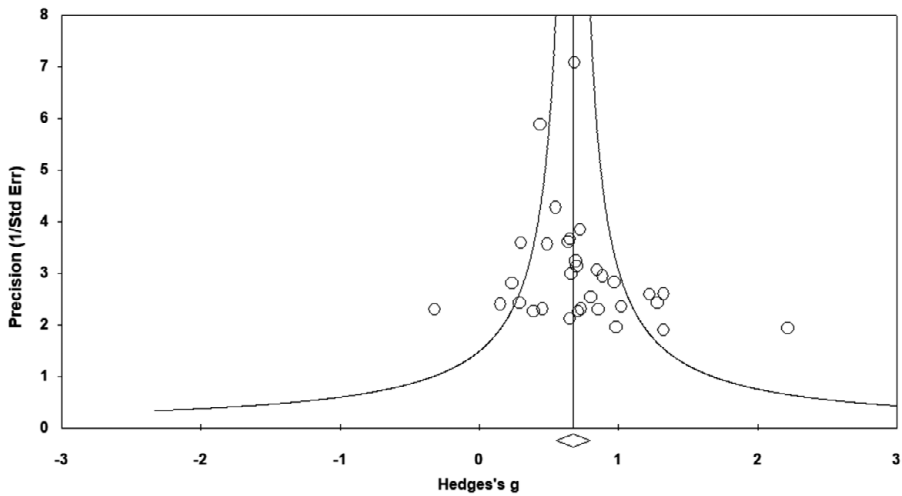

Figure 3. Funnel plot of perception effect size (mean treatment-control difference) by inverse standard error.

statistical significance but was negligible, g = –.08, SE = .04, 95% CI[–.16, .00], p = .058. A full moderator analysis was not executed due to the small sample size, but we conducted analyses for the second data (immediate vs. delayed posttest) on two relevant variables: retention intervals between posttests and delayed posttests (M = 2.3 months, range = .5 to 6) and proficiency (lower- vs. higher-level groups). Regarding L2 proficiency, due to the small sample size for each subcategory, the three proficiency levels (novice, beginner, and lower-intermediate) were combined to create a lower-level group; the two proficiency levels (upper-intermediate and advanced) were combined to create the higher-level group. Although we determined the cut-off point arbitrarily, we excluded the data from the intermediate group in this analysis to ensure the proficiency levels of the two larger groups would not overlap considerably. A meta-regression analysis showed no significant relationship between the interval and retention, b = –.004, SE = .025, 95% CI[–.054, .046], p = .882. The mean effect size was not significant for the higher-level group, k = 10, g = –.05, SE = .08, 95% CI[–.20, .10], p = .510, whereas it was small but significant for the lower-level group, k = 5, g = –.23, SE = .11, 95% CI [–.44, –.02], p = .028. Learners overall retained perception gains after HVPT, regardless of the retention interval between posttest and delayed posttest, but increased accuracy slightly declined over time for lower-proficiency learners.

RQ1c: Does improvement in L2 perception accuracy generalize to novel stimuli?

After one clear outlier was excluded (i.e., g = –6.14 in Ueda & Hashimoto, 2019), the aggregation of the effect sizes (i.e., the mean difference between posttest and generalized posttest scores) was conducted for each of the three categories of item and talker types: novel item and novel talker (k = 11), novel item and old talker (k = 14), and old item and novel talker (k = 5) (see Appendix 6 for detailed results and the forest plot). The mean effect size was not significant for the category of novel item and old talker, g = –.17, SE = .13, 95% CI [–.42, .08], p = .178. The mean effect size was significant but small for the category of novel item and novel talker, g = –.25, SE = .12, 95% CI[–.48, –.03], p = .028 and for the category of old item and novel talker, g = –.26, SE = .12, 95% CI[–.50, –.02], p = .031. The results of negligible to small differences in general indicate that improved perception accuracy generalized to novel stimuli, particularly when test items were produced by trained talkers.

RQ2: To what extent do learner-related and methodological variables moderate the effectiveness of HVPT for improving L2 perception accuracy?

Moderator analyses were conducted for 15 categorical and seven continuous variables with the pretest-posttest comparison data (k = 96). The analyses for three variables (response label, talker presentation, number of choices) were conducted based on studies involving identification training tasks (k = 82), given that these variables were of particular relevance to identification rather than discrimination training.[1]

Categorical Variables

The results of moderator analysis for 15 categorical variables are summarized in Table 2. No significant or noticeable differences between any of the sub-categories

[1]In discrimination training tasks, types of response labels were limited (e.g., same versus different), and the number of choices ranged from two to three (i.e., AX or ABX tasks). Discrimination tasks presented multiple voices in each trial; hence, the distinction between blocked and intermixed presentation was not easy to make.

Table 2. Moderator analyses for categorical variables (pretest-posttest comparison).

				95%CI			Q tests	
Variables	k	g	SE	LL	UL	p	Q	p
Learning context							1.22	.270
FL	78	.89	.05	.80	.99	.000		
SL	15	1.01	.10	.82	1.20	.000		
Proficiency							5.10	.531
Advanced	14	.81	.10	.62	.99	.000		
Upper-intermediate	4	1.22	.29	.65	1.78	.000		
Intermediate	7	1.00	.14	.73	1.27	.000		
Lower-intermediate	12	1.01	.13	.80	1.22	.000		
Beginner	14	1.01	.16	.70	1.33	.000		
Novice	16	.80	.10	.60	1.01	.000		
Target phone							28.33	<.001
Vowel	49	1.00	.07	.87	1.14	.000		
Obstruent	17	.69	.06	.57	.81	.000		
Sonorant	11	1.17	.11	.95	1.39	.000		
Syllable	3	1.41	.18	1.05	1.77	.000		
Tone	9	1.00	.20	.60	1.40	.000		
Inclusion of nonwords							.06	.809
Yes	38	.91	.07	.78	1.04	.000		
No	56	.93	.06	.81	1.05	.000		
Task type							15.54	<.001
Identification	82	.95	.05	.85	1.05	.000		
Discrimination	6	.57	.08	.41	.73	.000		
Response label							31.63	<.001
Keyword	31	1.03	.09	.86	1.20	.000		
Orthography	10	.90	.09	.72	1.08	.000		
Phonetic symbol	19	1.01	.11	.79	1.22	.000		
Visual image	5	.47	.07	.33	.61	.000		
Corrective feedback type							.12	.941
Target	47	.92	.06	.80	1.05	.000		
Combined	19	.96	.10	.77	1.15	.000		
Wrong	24	.92	.10	.73	1.11	.000		
Talker presentation							.84	.359
Blocked	40	.99	.07	.84	1.13	.000		
Intermixed	35	.89	.07	.75	1.04	.000		
Environment							1.11	.574
Laboratory	74	.95	.05	.85	1.05	.000		
Participant controlled	14	.83	.10	.63	1.03	.000		
Classroom	7	.87	.18	.52	1.21	.000		
Duration							4.50	.213
> 2 months	3	.75	.16	.44	1.05	.000		
1 to 2 months	28	.85	.07	.71	.99	.000		
1 week to 1 month	45	1.03	.08	.88	1.18	.000		
< 1 week	6	.81	.18	.46	1.15	.000		
Adaptive training							.27	.604
Yes	5	1.06	.28	.50	1.62	.000		
No	91	.91	.04	.82	1.00	.000		
Audiovisual input							1.17	.280
Yes	6	.82	.09	.65	.99	.000		
No	90	.93	.05	.83	1.02	.000		
Acoustic cue manipulation							.14	.708
Yes	3	1.01	.24	.54	1.48	.000		
No	93	.91	.05	.83	1.00	.000		
Test type							6.04	.014
Identification	85	1.08	.05	.97	1.18	.000		

(*Continued*)

Table 2. (*Continued*)

Variables	k	g	SE	95%CI LL	95%CI UL	p	Q tests Q	Q tests p
Discrimination	14	.77	.11	.54	.99	.000		
Item and talker type							6.00	.112
Old item : Old talker	25	.95	.08	.79	1.17	.000		
Old item : New talker	30	1.13	.13	.88	1.38	.000		
New item : Old talker	18	.76	.10	.56	.97	.000		
New item : New talker	51	1.03	.09	.86	1.20	.000		

Note: k = number of studies. CI = confidence interval.

($g_{difference}$ < .40 or defined as a small effect according to Plonsky & Oswald, 2014) were found for the following variables: learning context, proficiency, inclusion of nonwords, corrective feedback type, talker presentation, environment, duration, adaptive training, audiovisual input, acoustic cue manipulation, and item and talker type. The variables found as significant moderators include target phone (p < .001), task type (p < .001), response label (p < .001), and test type (p = .014).

Regarding the variable of target phone, the largest effect size was observed for syllable structure (g = 1.41), followed by sonorant[2] (g = 1.17), vowel (g = 1.00), tone (g = 1.00), and obstruent (g = .69). Regarding the variable of response label, a lower training effect was observed for studies using visual images (g = .47) compared to studies using other types of response labels (g = 1.03 for keywords, .90 for orthographies, and 1.01 for phonetic symbols). As for the variable of task type, a larger effect was found for the identification training task (g = .95) compared to the discrimination training task (g = .57). Regarding the variable of test type, a larger effect was observed when perception accuracy was measured with the identification test (g = 1.08) than the discrimination test (g = .77). A further examination of the data revealed that when learners were trained on the identification task, a larger effect was found for the identification test (k = 73, g = 1.11, SE = .06, 95% CI[.99, 1.23], p < .001) than when perception accuracy was trained on the discrimination task and measured with the identification test (k = 6, g = .67, SE = .11, 95% CI[.46, .88], p < .001). The effect sizes were comparable for the discrimination test performance between identification trainees (k = 12, g = .80, SE = .13, 95% CI[.54, 1.06], p < .001) and discrimination trainees (k = 1, g = .84, SE = .32, 95% CI[.21, 1.48], p = .009). However, the finding for the discrimination test needs to be interpreted with caution due to the small sample size.

Continuous variables

Meta-regression analyses were conducted for two variables of learner features (AOL and AOT), two variables of training intensity (training time per session and total training time), and three variables of training format (number of talkers, phonetic contexts, and choices).

Learner features

Regarding learner-related variables, a multiple meta-regression analysis with AOL and AOT (r = .558, 95% CI[.280, .749], p < .001) as predictor variables (k = 34) showed that

[2]All studies coded as sonorants except Yang et al. (2021) (i.e., nasal codas) examined liquids (/l/ and /r/).

AOT was significantly and negatively associated with perception improvement (b = –.039, SE = .018, 95% CI[–.073, –.005], p = .027), whereas AOL was not a significant predictor of perception improvement (b = .031, SE = .021, 95% CI[–.010, .071], p = .140). The result that AOT was a significant variable while the effect of AOL was constant indicates that the length of L2 learning may relate to perception gain. To confirm this finding, the length of L2 learning was calculated by subtracting AOL from AOT, and an additional analysis with this new variable as a predictor was conducted. The result was consistent with the previous result, showing a significant and negative relationship between the length of learning and perception improvement, b = –.036, SE = .017, 95% CI[–.069, –.004], p = .029 (for the scatterplot of the two variables, see Figure 4).

Training intensity

Two separate meta-regression analyses were conducted for training time per session and total training time (i.e., the product of training time and number of sessions). Training time per session (k = 79) was not significantly associated with perception gain, b = .0027, SE = .0026, 95% CI[–.0024, .0077], p = .298. There was no significant association between total training time (k = 80) and perception gain, b = .0003, SE = .0002, 95% CI[–.0002, .0007], p = .228. However, Figure 5 shows that the relationship between total training time and perception gain does not seem linear, with the gain after 400 minutes or more of training leveling off. A follow-up analysis focusing exclusively on studies involving the total amount of training time with less than 400 minutes (k = 58) revealed a significant and positive association between total training time and perception improvement, b = .0034, SE = .0008, 95% CI[.0019, .0050], p < .001. The results indicate that the increased training time continued to have a positive effect on perception accuracy until it reached around 400 minutes of training.

Training format

A multiple meta-regression analysis was conducted with the three moderator variables (number of talkers, contexts, and choices) as predictors. Number of target sounds was included as a covariate for each of the three regression models because the three

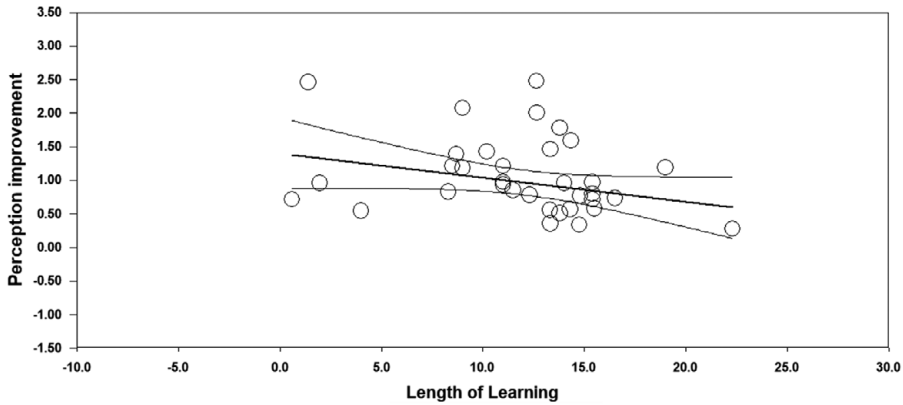

Figure 4. The relationship between the length of L2 learning and perception improvement with a 95% confidence interval.

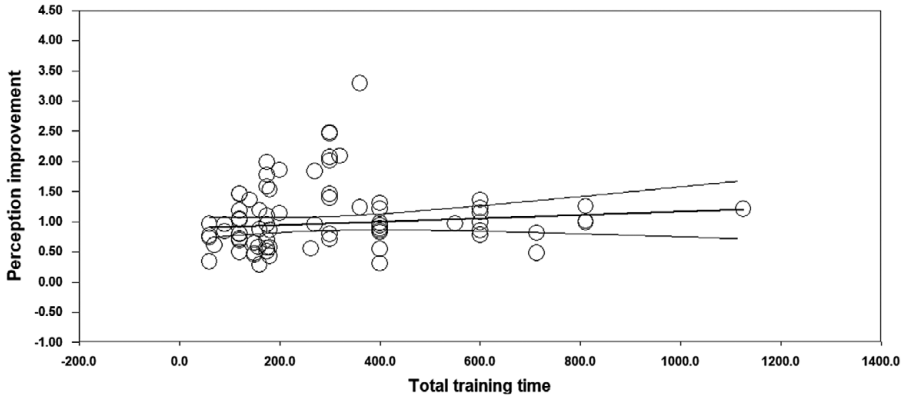

Figure 5. The relationship between total training time and perception improvement with a 95% confidence interval.

variables appear to be biased and correlated with the number of sounds targeted in the studies: the correlation between number of sounds and number of talkers (r = –.31, p = .002), phonetic contexts (r = –.53, p < .001), and response choices (r = .47, p < .001). The meta-regression analysis showed that none of the variables was statistically significant: phonetic context (k = 88, b = .0012, SE = .0016, 95% CI[–.0020, .0044], p = .459), number of response choices (k = 79, b = .0156, SE = .0203, 95% CI[–.0241, .0553], p = .442), and number of talkers (k = 93, b = –.0095, SE = .0087, 95% CI[–.0265, .0076], p = .278).

The initial analysis of several talkers, as presented in Figure 6, appears to show a tendency that an increased number of talkers (from 10 up to 30) lowered perception gains. We revisited the data on several talkers and reexamined their effect on perception gain while considering the potential influence of L2 proficiency. In this follow-up analysis, we focused on the data within the range from 2 to 10 talkers (k = 88), given that this range of data accounted for 91.7% of the entire data set (i.e., 88 out of 96 studies), while the data of 10 and above (up to 30 talkers) was represented by only eight studies. We conducted a meta-regression analysis with the number of talkers as the predictor for the two proficiency groups separately: the higher-level group, consisting of upper-

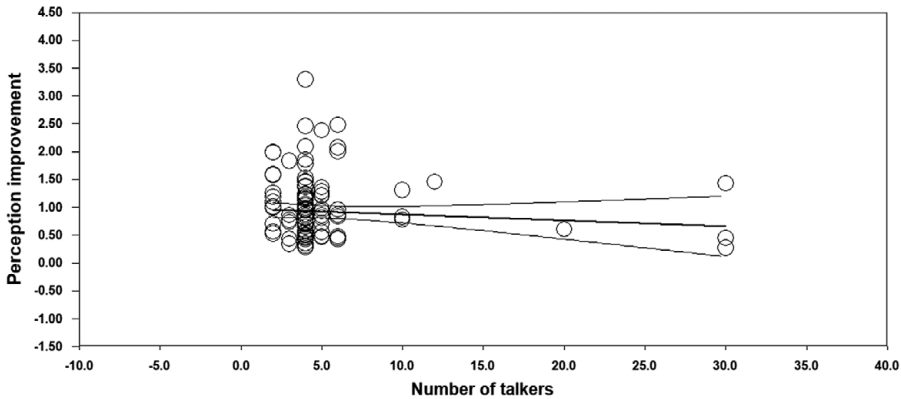

Figure 6. The relationship between the number of talkers and perception improvement with a 95% confidence interval.

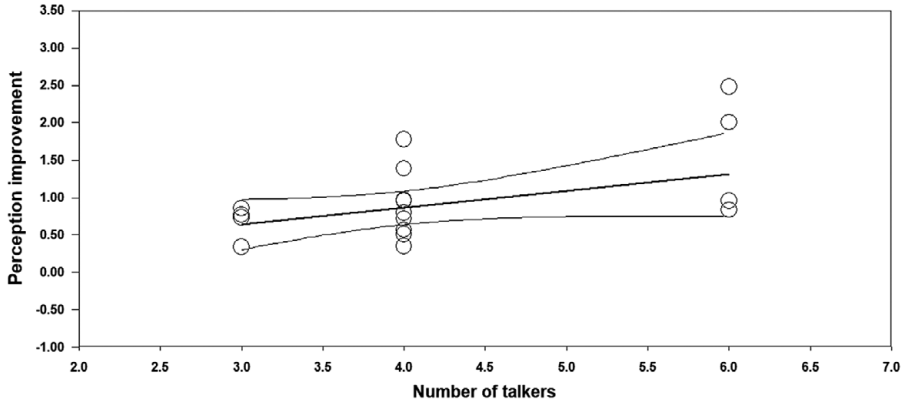

Figure 7. The relationship between number of talkers and perception improvement for the higher-level group with a 95% confidence interval.

intermediate and advanced learners (k = 17), and the lower-level group, consisting of novice, beginner, and lower-intermediate learners (k = 39).

The additional analysis showed that for the higher-level group (Figure 7) the increase in the number of talkers significantly improved perception gain, b = .22, SE = .10, 95% CI[.02, .43], p = .032, while for the lower-level group (Figure 8) no significant effect of number of talkers was observed, b = –.06, SE = .07, 95% CI[–.20, .09], p = .446. These results and the summary of the effect sizes (see Table 3) demonstrated that the higher-level group benefited from the increased number of talkers up to six talkers with a large effect (g = 1.44), while the lower-level group did not appear to receive additional benefits from the increase in the number of talkers.

RQ3: To what extent does L2 perception accuracy increase in untrained learner controls?

Regarding methodological concerns in the field of speech training research, the degree to which control groups taking tests without completing training sessions or receiving

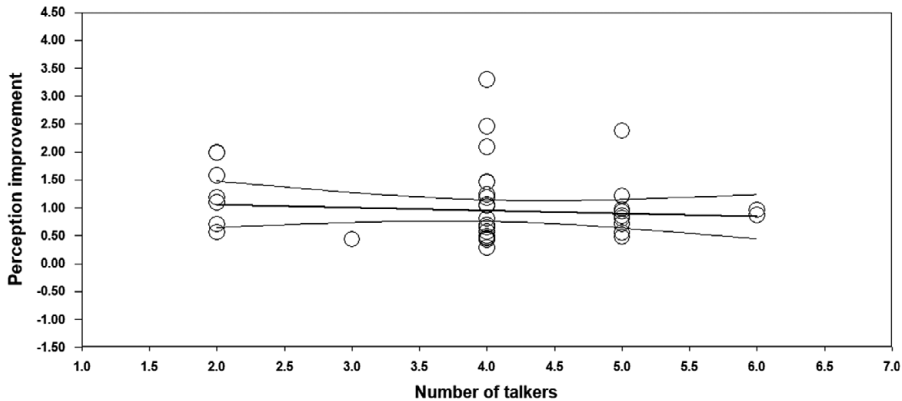

Figure 8. The relationship between number of talkers and perception improvement for the lower-level group with a 95% confidence interval.

Table 3. Summary of the relationship between the number of talkers and perception gain for higher-level and lower-level groups

	Higher-level group			Lower-level group		
Number of talkers	k	g	95% CI	k	g	95% CI
2	0	–	–	7	1.17	[.79, 1.54]
3	4	.66	[.43, .90]	1	.43	[.03, .84]
4	9	.85	[.60, 1.11]	19	.96	[.71, 1.21]
5	0	–	–	10	.89	[.68, 1.10]
6	4	1.44	[.73, 2.15]	2	.90	[.62, 1.19]
>10	1	.78	[.50, 1.06]	3	.73	[.79, 1.54]

Note: k = number of studies. CI = confidence interval.

training with nontarget sounds was examined (see Appendix 6 for detailed results and the forest plot). Based on 28 out of 31 studies reporting perception gains in percentage, the unweighted mean gain for the control groups was 2.7% (SD = 3.8, 95% CI [2.3, 3.0]). The aggregation of the 31 effect sizes showed that the mean effect size was significant but small, g = .19, SE = .07, 95% CI [.05, .33], p = .009. Because learning context (FL vs. SL) was expected to pertain to the amount of exposure to L2 input learners receive outside the classroom, a further analysis of the difference between FL (k = 27) and SL (k = 4) was conducted. The mean effect size was significant and small for FL, g = .17, SE = .08, 95% CI [.01, .32], p = .034. The mean effect size for SL was also significant and twice as large as the effect size for FL, g = .37, SE = .15, 95% CI [.09, .70], p = .011. Lastly, to examine the effect of familiarity with training task procedure, we compared a test-only group that took only tests without completing any training (k = 26) with another control group who took tests and completed training on nontargeted phones (k = 5). A nonsignificant but slightly larger gain was found for the control group with training on nontargeted sounds, g = .28, SE = .24, 95% CI [–.20, .75], p = .251, compared to the test-only group, g = .16, SE = .07, 95% CI [.02, .31], p = .022.

Discussion

RQ1: The effectiveness of HVPT for perception improvement, retention, and generalization

Our meta-analysis results show a medium-to-large effect of HVPT on L2 perceptual learning (g = .92 for within-participant design; g = .67 for between-participant design). Note, however, that the effect size for the within-participant design, after adjusted for potential publication bias, dropped from .92 to .71 (i.e., medium effect). The overall rate of L2 perception improvement was 14.12% for trained stimuli (i.e., old item-old talker) and 12.96% for untrained stimuli (i.e., new item-new talker). The moderator analysis for item and talker types (i.e., partial generalization) showed a nonsignificant variation among different types of stimuli, indicating the generalization of learning to untrained stimuli and talkers. This evidence corroborates the effectiveness of HVPT for learning stimulus-general as well as stimulus-specific structures. Compared to the mean effect size (g = .28 for immediate posttest) reported in Zhang et al. (2021), the effect size observed in this study was much larger (e.g., g = .67). This is arguably because the larger effect of HVPT in the current meta-analysis was the result of the overall training efficacy with positive effects from three training features combined (i.e., talker variability, context variability, and corrective feedback), while in Zhang et al.'s study, the

small effect mirrored that of only one of the contributing factors (i.e., talker variability). Given the recent meta-analysis of HVPT focusing on L2 speech production (Uchihara et al., 2024), perception gains observed in this study are considered larger than production gains (10.56% for trained items, 4.5% for untrained items; g = .49 for within-participant design, g = .66 for between-participant design). These findings dovetail with the perception-first view of L2 speech development (Flege, 1995a).

Regarding the results of retention, delayed posttest accuracy significantly outperformed pretest accuracy (g = .98 or .73 after adjusted for publication bias). There was also a negligible and nonsignificant difference between posttest and delayed posttest performance (g = –.08) regardless of the retention interval (i.e., .5 to 6 months). These findings suggest that perceptual learning as a result of HVPT is likely to be retained over time, especially for higher proficiency learners (g = –.05), noting that the retention of learning for lower proficiency learners decayed to a small degree (g = –.23, p = .028). Regarding complete generalization, there was a negligible difference in perception accuracy between posttest and generalization posttest performance based on stimuli of novel items (g = –.17, p = .131), or if significant, the differences were small for trained stimuli produced by novel talkers (g = –.26, p = .031) and novel stimuli produced by novel talkers (g = –.25, p = .028). While this result needs to be considered in later research to further optimize the effectiveness of HVPT for acquiring L2 categories robust to talker variations, it may be safe to say that HVPT works in general for learning L2 phonetic categories beyond item- or talker-specific acoustic features.

These findings confirm the effectiveness of HVPT for perceptual learning, retention, and generalization, highlighting the practical usefulness of this training technique for L2 speech instruction. The rate of learning (i.e., 12.96 to 14.12%) found in this meta-analysis was also roughly consistent with that expected by L2 speech researchers (e.g., 10 to 15%, Flege, 1995b; Iverson et al., 2012). However, a significant heterogeneity in the effect size across studies indicated that the actual perception gains varied considerably across individual studies. This result motivated us to explore variables that significantly moderated the effectiveness of HVPT.

RQ2: Variables moderating the effectiveness of HVPT

We analyzed 22 moderator variables related to learner profiles, training features, and testing features. Most were not identified as significant moderators, or did not have a sizable impact on perceptual training: i.e., learning context, proficiency, inclusion of nonwords, corrective feedback type, talker presentation, environment, duration, audio-visual, acoustic cue manipulation, item and talker type, AOL, training time per session, phonetic context, and number of response choices. On the other hand, seven variables were found to have a major impact on the effectiveness of perceptual training: AOT while AOL was controlled (or length of L2 learning), type of training task, type of testing task, total training time, target phone, response label, and number of talkers. In what follows, we will discuss some of the key findings that deserve further attention in L2 speech research and instruction.

Learner features

The lack of significant variation for L2 proficiency and learning context indicates that HVPT may be equally beneficial for learners with different proficiency levels in both input-limited and input-rich contexts. These findings suggest that L2 experience may not be a critical factor that affects perceptual learning (Iverson et al., 2012; Wong, 2014).

However, a closer examination of the proficiency data suggests that the effect size was slightly smaller for the advanced group (g = .81) and novice learners (g = .80) compared to other groups (g ≧ 1.00). As for novice learners, perhaps various factors (e.g., lack of motivation to learn the L2) might influence their perceptual learning. Besides the moderator analyses, also relevant to the effect of L2 experience was the finding of a significant correlation between pretraining perception accuracy and perception improvement (r = –.52, p < .001). The negative correlation indicates that learners with higher perceptional abilities tend to benefit less from HVPT. The possibility that this negative correlation was the artifact of a ceiling effect for some primary studies is less likely because all studies except one (M = 89%, Lengeris & Nicolaidis, 2014) did not exceed 85% points, a ceiling upper cut-off point commonly adopted in HVPT research (Brekelmans et al., 2022; Iverson et al., 2012): $M_{pretest}$ = 63.7%, SD = 9.3%, range = 36.4 to 82.5%. Additional support for the negative impact of L2 experience comes from the finding that the length of L2 learning (i.e., the value computed by subtracting AOL from AOT) was negatively associated with perception gains. These findings imply that learners with more L2 learning experience tended to benefit less from HVPT. However, this result needs to be interpreted with caution given that our rough estimate of the learning length cannot be equated to the actual amount of L2 input or may be confounded by other learner-internal or -external variables (e.g., motivation, learning context).

One possible reason for these findings is that learners who have higher pretest performance may have limited room left for improvement, regardless of whether learners reached a ceiling performance. Another reason may relate to the established nonnative knowledge of L2 sounds and its negative influence on perception improvement. Given that the majority of studies included in this meta-analysis were conducted in foreign language environments (k = 79), learners who have spent longer time learning the L2 might have developed L1-specific manners of perceiving L2 sounds (Tyler, 2019). Such nonnative phonetic knowledge might interfere with the process of fundamentally changing category representations (e.g., cue weightings) and slow down the acquisition of nativelike phonetic knowledge. These findings may explain how L2 experience and proficiency impact HVPT, and why advanced learners may show less benefit.

Training and testing features

In this section, we discuss the impact of seven moderator variables: type of training task, type of testing task, total training time, target phones, response labels, number of talkers, and response labels.

First, a much larger training effect was found for identification (ID) tasks (g = .95) than for discrimination (DIS) tasks (g = .57). This finding supports the claim that ID training is more appropriate for improving L2 perception accuracy than DIS training (e.g., Carlet & Cebrian, 2019; Carlet, 2017). However, since the majority of studies in the current meta-analysis used ID training tasks (k = 82), with only six studies employing DIS tasks, this finding needs to be further explored. The overlap between training and testing tasks for ID in the majority of studies might indicate that learning effects are optimized when the testing and training formats are matched (Morris, Bransford, & Franks, 1977). However, we found a relatively larger effect for the training-testing matched condition for ID (g = 1.11 for ID training–ID testing) than for DIS (g = .84 for DIS training–DIS testing). This finding seems to confirm the superiority of ID training for L2 perception improvement. Yet, the effect size for the mismatched condition for ID

(g = .67 for DIS training–ID testing) was much smaller than the matched condition (g = 1.11 for ID training–ID testing). This is consistent with the view that ID and DIS tasks tap into different aspects of L2 phonetic knowledge (Carlet & Cebrian, 2022; de Oliveira, 2020; Iverson et al., 2012; Leong et al., 2018). Enhancing a lower level and sensory mode of processing through DIS tasks may have a positive but restricted effect on identification performance, requiring a higher-level phonological encoding of L2 speech (but see Cebrian et al., 2024 for a review of differential effectiveness of auditory and categorical discrimination tasks).

Regarding training intensity, the total training time did not significantly predict perception improvement. This finding contrasts with the results of Zhang et al.'s (2021) meta-analysis demonstrating that a longer length of training led to greater learning from exposure to talker variability input. The conflicting results may be attributed to some methodological differences. First, Zhang et al. (2021) focused exclusively on the contrast between multitalker and single-talker conditions, whereas the current study investigated the overall gain from pretest to posttest performance. Second, the range in the training length in Zhang et al. (2021) was restricted to 60 to 480 minutes, whereas the current meta-analysis contained 14 studies involving the total training time beyond 480 minutes (i.e., 500 to 1,125 minutes). A follow-up analysis revealed that the total training time up to fewer than 400 minutes was a significant and positive predictor of perception improvement, but beyond 400 minutes, perception gain appeared to level off up to 1,125 minutes. These findings suggest that the duration of training necessary to reach maximum improvement may be around 400 (or up to 480) minutes. Longer training time is less likely to bring about further improvement for several reasons such as learner fatigue or hitting a ceiling in improvement (Thomson, 2018).

Regarding the target phones, one notable finding was that prosodic features of tones and syllable structures enjoyed large training effects (g = 1.00 for tones and g = 1.41 for syllable structure). These findings support the efficacy of HVPT for improving not only segmental perception accuracy but also suprasegmental accuracy, especially for learning phonotactic constraints on syllable structures (e.g., Huensch & Tremblay, 2015). Another notable finding was that a larger effect was observed for sonorants (mostly /l/ and /r/) (g = 1.17) than for obstruents (g = .69), given the widely perceived difficulty in acquiring sonorants (see Sakai & Moorman, 2018 for production data). However, these findings need to be interpreted with caution because studies in the current meta-analysis tended to spare more training time for sonorants ($M_{\text{total training time}}$ = 430 min) than for obstruents ($M_{\text{total training time}}$ = 227 min). Perhaps researchers are aware of the difficulty in learning sonorants (e.g., /r/ vs. /l/) and tended to spare more time for training on such phonemes (Bradlow, 2008), whereas obstruents are considered to be relatively easier to perceive and learn with a few training sessions (Carlet & Cebrian, 2022).

With respect to response labels, a much lower training effect was observed for studies using visual images (g = .47) compared to studies using other types of response labels (g ≧ .90 for keywords, orthographies, and phonetic symbols). Visual and pictorial labels were used in earlier studies (e.g., Giannakopoulou, Brown, Clayards, & Wonnacott, 2017) to prevent associations between training stimuli and potentially faulty sound-spelling knowledge learners acquired previously (Fouz-González & Mompean, 2021). However, the use of these nonorthographic labels might induce a cognitively demanding process in L2 speech perception, which might obscure the efficacy of perceptual training. Response to orthographic labels only requires listeners to attend to the auditory form of words (e.g., the initial phoneme of *hat* and *cat*), whereas identifying the right nonorthographic labels involves processing both the auditory form (e.g., /h/ vs. /k/) and the knowledge of form-meaning mapping

(e.g., the link between /hæt/ and a picture conveying a shaped covering for the head). According to the account of limited cognitive capacity and processing allocation (Barcroft, 2002), dispersing attentional resources to multiple aspects of lexical knowledge may reduce the amount of attention allocated to the encoding of target forms and thus lower the overall training effects (for a description of learning multiple aspects of spoken lexical knowledge, see Uchihara, Webb, Saito, & Trofimovich, 2022).

Regarding the number of talkers, no significant differences were found within the range of two and 30 talkers, with a tendency showing that training effects became slightly smaller within the range of 20 and 30 talkers. Our follow-up analysis, however, revealed that higher-proficiency learners benefit from an increased number of talkers (i.e., three to six talkers, g = .66 → 1.44), while the lower-proficiency learners consistently experienced a benefit irrespective of a widely varying number of talkers (e.g., g = 1.17 for two talkers, .96 for four talkers, .89 for five talkers). In response to the call for determining the optimal number of talkers for the best practice of HVPT (Thomson, 2018), the current findings suggest that it is somewhere in the vicinity of six talkers, particularly for higher proficiency learners. However, this result should be interpreted with caution, considering the complex interplay of other variables that may have been masked in the current meta-analysis. For instance, comparing two hypothetical studies, Study A using three talkers, and Study B using five talkers, might indicate that the condition of three talkers is better than that of five. However, it might simply be the case that Study A's three talkers includes one which, by coincidence, is an exemplary, easy-to-learn-from model. In contrast, Study B's five talkers may coincidentally all be relatively poor models for a particular group of learners. Thus, these hypothetical findings would make it difficult to tease apart the effect of talker variability from that of talker quality, failing to yield the advantage associated with talker variability for Study B. The definite answer to the question of the optimal number of talkers awaits more primary studies comparing varying numbers of talkers, while also explicitly controlling the influence of talker quality.

RQ3: The evaluation of perception gains for control groups

Because exploration of methodological features in primary studies is one of the responsibilities of meta-analysis, we calculated the estimated gain for control groups, an indication of the magnitude of effects resulting from training-irrelevant variables such as testing effects, out-of-experiment exposure to L2 input, and training task familiarity. Results showed that the overall gain for the control groups was 2.7%, indicating that learners not trained on target sounds improved perception accuracy by around 3%. The aggregated effect size found was significant (g = .19) and appeared to be smaller than that of L2 pronunciation gain (d = .31) for test-only control groups (Saito & Plonsky, 2019). The effect size was larger than complex forms (d = .04) and smaller than simple forms (d = .28) reported in L2 grammar learning (Spada & Tomita, 2010). Based on these findings, it is important for future studies, especially when they do not involve control groups, to interpret perception gains with caution; for example, if 15% gain is found, the true gain may go down to around 12.3% (= 15–2.7%). Furthermore, despite the small overall effect for the control groups, the effect size appeared to increase up to g = .37 when studies were conducted in L2 immersion contexts (SL), compared to a much smaller effect (g = .17) in L2 input-limited contexts (FL). Regarding the influence of training task familiarity, control groups taking tests and completing training on non-targeted sounds produced a relatively larger effect (g = .28), compared to control groups

taking tests without receiving any training (g = .16). These findings suggest that exposure to L2 input (SL vs. FL) and task familiarity (with nontarget training vs. without nontarget training) may increase the gain for control groups (Cebrian et al., 2024). Although individual effects are not considered large, accumulation of the effects (e.g., experiments conducted in SL contexts involving nontarget training) might have a substantial impact on the extent to which true HVPT effects are confounded. Calculating the actual percentage gain for such a condition was not feasible due to the lack of relevant studies (i.e., SL + nontarget training), but an even larger gain would be expected to emerge.

Conclusion, implications, and future directions

The present study conducted a meta-analysis on 30 years of HVPT studies with the aim of determining the effectiveness of HVPT for L2 perceptual learning and identifying key moderator variables that influence the training effects. This study confirmed the effectiveness of the HVPT technique for improving perception accuracy of L2 sounds, long-term retention of perceptual learning, and generalization to untrained stimuli. The moderator analyses identified several key variables that enhance (or hinder) the effectiveness of HVPT, including the length of L2 learning, target phones, response labels, type of training task, type of testing task, total training time, and number of talkers. To close, we discuss practical implications for optimizing HVPT in L2 instructional settings, methodological suggestions for future research, and limitations of this study that can be expanded upon to point to future directions in HVPT research.

Pedagogical implications

The findings of the current meta-analysis provide compelling evidence showing that HVPT can benefit the learning and teaching of L2 speech sounds. One possible way to integrate HVPT into L2 classrooms is to utilize it in a supplemental manner to support ongoing classroom instructions, especially in teaching individual sounds that are challenging to acquire (e.g., English /l/ and /r/ for L1 Japanese learners). However, it is important to note that HVPT, considered a decontextualized form-focused practice, should not be a primary focus in the classroom. Rather, such practice should be balanced with meaning-focused activities (e.g., task-based teaching; Mora & Levkina, 2017), given that the ultimate goal of L2 instruction is to develop the ability to comprehend (and produce) L2 speech in real-life communicative contexts (Saito & Plonsky, 2019). Perhaps teachers interested in using HVPT should first prioritize communicatively important segmental features that may have a major impact on overall comprehensibility (Suzukida & Saito, 2021). They can also conduct initial needs assessments with their learners using diagnostic evaluations (Firth, 1992; Isbell, 2021) and then strategically direct their students to use existing training platforms such as English Accent Coach (Thomson, 2012a) or gamified versions of HVPT (Saito, Hanzawa, Petrova, Kachlicka, Suzukida, & Tierney, 2022; Thomson, 2024) to further enhance students' perceptual abilities while saving valuable classroom time for communicative instruction.

With these caveats in mind, the results of our analysis show key features that should be considered for material developers and teachers designing speech training programs for their students in terms of training task, response labels, total training time, and number of talkers. First, it should be noted that identification training is more likely to enhance the amount of perceptual learning compared to discrimination training. When

adopting identification training tasks, teachers need to consider the type of response choices. Providing new referents (e.g., visual images) as response choices may not be a recommended practice when the primary focus of L2 instruction is on improving perception of L2 sounds in terms of the knowledge of auditory forms, not the knowledge of form-referent mapping. On the bright side, our findings (i.e., small but significant effect for visual images, g = .47) imply that perceptual learning occurs when learners' attention is drawn to not only processing the auditory forms but also mapping the forms to referents, the situation reflecting real-life L2 spoken word learning. Thus, advice against the use of visual images is not always appropriate in contexts, for example, where the goal of L2 instruction with HVPT is to learn a new word's meaning as well as improve accuracy at perceiving a given L2 sound that appears in the word. Second, longer training time does not necessarily lead to greater return for perceptual learning. The substantial improvement may not be expected after the point of around 400 minutes (or 6.7 hours) of training. Lastly, we suggest approximately six as the optimal number of talkers, especially for higher proficiency learners. For lower proficiency learners, the number of talkers may not exert a measurable impact on perception gains if the range in the number of talkers lies somewhere between two and six. While we cannot confidently say that these numbers should be interpreted as the gold standard, we do feel that evidence to date reasonably supports this recommendation as the starting point in developing a perceptual training program.

Methodological implications

The current meta-analysis provides methodological recommendations for future studies. First, virtually no studies except one (Qian, 2018) reported perception test reliabilities. Besides encouragement of reporting basic statistical information (e.g., means and standard deviations for pretest and posttest perception scores), we urge researchers to also report the pretest and posttest reliabilities and evaluate the impact of test reliability on their results. Second, sample size needs to be increased in future HVPT research (for the same call, see also Sakai & Moorman, 2018). We found that the average sample size in this meta-analysis was 17.2 and there has been no clear increase in the number of participants since Logan et al. (1991). Given the great amount of time and cost required for the implementation of HVPT experiments, researchers may find online experiment builders such as Gorilla (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020) or online training platforms (e.g., English Accent Coach, Thomson, 2012a) useful. Third, future studies should consider whether participants receive exposure to target stimuli during the pretest and distinguish "novel" from "new" stimuli in making claims about generalization to untrained stimuli. The different findings regarding generalization (e.g., nonsignificant between-category differences for partial generalization, while some significant differences for complete generalization involving novel talkers) indicate that this is an important distinction. To determine the true or complete generalization of perceptual learning, it is important to explore whether HVPT is effective in generalizing to novel stimuli and if not, a further investigation needs to focus on how it can be improved, for example, with modifications to conventional HVPT (e.g., adaptive training, audiovisual input). Finally, future studies should consider the potential impact of training-irrelevant variables (e.g., testing effects, out-of-experiment exposure to L2 input, and training task familiarity). Particularly, when studies aim to track perceptual learning over an extended period of time in input-rich, L2 immersion contexts, it is ideal to incorporate a test-only control group.

In fact, the lack of attention to the impact of training-irrelevant variables can be attested by the observation that the treatment-control comparison data was reported and available for analysis in only 30 out of 79 papers (38%). As with the increase in the number of studies integrating the data from control conditions, a future meta-analysis would also reveal a more accurate picture of HVPT efficacy and further clarify variables that moderate the training effects.

Limitations and future directions

We note several limitations of our meta-analysis and suggest future directions that can be built upon them. First, nonsignificant results of moderator analysis should be interpreted with caution. There were several variables not significantly predictive of perceptual learning (e.g., audiovisual input, type of corrective feedback, and the order of talker presentation). However, these results should not be interpreted as direct evidence against the effectiveness of such variables for perceptual learning. For example, the studies investigating audiovisual input were not randomly assigned to levels of "audio-visual input" versus "audio-only input," and casual inferences should not be drawn based merely on the current findings. If the primary purpose of a meta-analysis is to examine the effect of audiovisual input, a future meta-analysis should focus exclusively on studies contrasting audiovisual input with audio-only input conditions. Second, the results of the current meta-analysis may be influenced by the presence of potential publication bias. For instance, the mean effect size of pretest-posttest contrast was initially considered large (g = .92) but turned out to be medium (g = .71) after corrected for publication bias. Consequently, the results of moderator analyses using the same data may have been influenced by the publication bias. This issue is inseparable from the problem of small sample size in the domain of HVPT research, pointing to the possibility that studies with smaller sample size (and larger standard errors) reporting larger training effects are more likely to be accepted for publication than those reporting smaller training effects. While reiterating the importance of increasing sample size in HVPT research, we suggest that all stakeholders, including researchers, meta-analysts, and journal editors, may need to seriously consider the issues of publication bias in the field of L2 speech training research.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0272263125100879.

Acknowledgments. We gratefully acknowledge insightful comments from anonymous reviewers and Editor Luke Plonsky on earlier versions of the manuscript. We also thank the following researchers who kindly provided information necessary for the current meta-analysis to be completed: Angélica Carlet, Juli Cebrian, Payam Ghaffarvand Mokari, Hyosung Hwang, Na-Young Ryu, Ruining Yang, and Angelos Lengeris.

Competing interest. The authors declare none.

References

Aliaga-Garcia, C. (2017). *The effect of auditory and articulatory phonetic training on the perception and production of L2 vowels by Catalan-Spanish learners of English* [Doctoral dissertation, Universitat de Barcelona].

Antoniou, M., & Wong, P. (2015). Poor phonetic perceivers are affected by cognitive load when resolving talker variability. *The Journal of the Acoustical Society of America*, *138*(2), 571–574. http://doi.org/10.1121/1.4923362

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Barcroft, J. (2002). Semantic and structural elaboration in L2 lexical acquisition. *Language Learning*, *52*(2), 323–363. https://doi.org/10.1111/0023-8333.00186

Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition*, *27*, 387–414. https://doi.org/10.1017/S0272263105050175

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.

Borenstein, M. (2022). In a meta-analysis, the I-squared statistic does not tell us how much the effect size varies. *Journal of Clinical Epidemiology*, *152*, 281–284. https://doi.org/10.1016/j.clinepi.2022.10.003

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis*. Wiley.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2022). Comprehensive meta-analysis (Version 4.0) [Computer software]. Biostat. https://www.meta-analysis.com/

Bradlow, A. R. (2008). Training non-native language sound patterns: Lessons from training Japanese adults on the English /ɹ/-/l/ contrast. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 287–308). Benjamins.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics*, *61*(5), 977–985. https://doi.org/10.3758/BF03206911

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, *101*(4), 2299–2310. https://doi.org/10.1121/1.418276

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

Burnham, K. R. (2013–2014). Phonetic training in the foreign language curriculum. *Applied Language Learning*, 23–24, 63–74.

Bulgarelli, F., & Weiss, D. J. (2021). Desirable difficulties in language learning? How talker variability impacts artificial grammar learning. *Language Learning*, *71*(4), 1085–1121. https://doi.org/10.1111/lang.12464

Carlet, A., & Cebrian, J. (2019). Assessing the effect of perceptual training on L2 vowel identification, generalization and long-term effects. In A. M. Nyvad, M. Hejná, A. Højen, A. Bothe Jespersen, & M. Hjortshøj Sørensen (Eds), *A sound approach to language matters—In honor of Ocke-Schwen Bohn* (pp. 91–119). Aarhus University.

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

Carlet, A. (2017). *L2 perception and production of English consonants and vowels by Catalan speakers: The effects of attention and training task in a cross-training study* [Doctoral dissertation, Universitat Autònoma de Barcelona].

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*, 1–25. https://doi.org/10.1017/S0272263124000408

Del Re, A. C., & Hoyt, W. T. (2010). *MAd: Meta-analysis with mean differences* (R Package Version 0.8) [Computer software]. https://cran.r-project.org/web/packages/MAd/index.html

de Oliveira, D. M. (2020). *Auditory selective attention and performance in high variability phonetic training: The perception of Portuguese stops by Chinese L2 learners* [Doctoral dissertation, Universidade do Minho].

Firth, S. (1992). Pronunciation syllabus design: A question of focus. In P. Avery & S. Ehrlich (Eds.), *Teaching American English pronunciation* (pp. 173–183). Oxford University Press.

Flege, J. E. (1995a). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 229–273). York Press.

Flege, J. E. (1995b). Two procedures for training a novel second language phonetic contrast. *Applied Psycholinguistics*, *16*(4), 425–442. https://doi.org/10.1017/S0142716400066029

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

Fuhrmeister, P., & Myers, E. B. (2020). Desirable and undesirable difficulties: Influences of variability, training schedule, and aptitude on nonnative phonetic learning. *Attention, Perception, & Psychophysics*, *82*(4), 2049–2065. https://doi.org/10.3758/s13414-019-01925-y

Georgiou, G. P. (2021). Effects of phonetic training on the discrimination of second language sounds by learners with naturalistic access to the second language. *Journal of Psycholinguistic Research*, *50*(3), 707–721. https://doi.org/10.1007/s10936-021-09774-3

Giannakopoulou, A., Brown, H., Clayards, M., & Wonnacott, E. (2017). High or low? Comparing high and low-variability phonetic training in adult and child second language learners. *PeerJ*, *5*, e3209. https://doi.org/10.7717/peerj.3209

Giannakopoulou, A., Uther, M., & Ylinen, S. (2013). Enhanced plasticity in spoken language acquisition for child learners: Evidence from phonetic training studies in child and adult learners of English. *Child Language Teaching and Therapy*, *29*(2), 201–218. https://doi.org/10.1177/0265659012467473

Hardison, D. M. (2003). Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics*, *24*(4), 495–522. https://doi.org/10.1017/S0142716403000250

Herd, W., Jongman, A., & Sereno, J. (2013). Perceptual and production training of intervocalic /d, ɾ, ɹ/ in American English learners of Spanish. *The Journal of the Acoustical Society of America*, *133*(6), 4247–4255. https://doi.org/10.1121/1.4802902

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, *327*, 557–560. https://doi.org/10.1136/bmj.327.7414.557

Hirata, Y., & Kelly, S. D. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *Journal of Speech, Language, and Hearing Research*, *53*(2), 298–310. https://doi.org/10.1044/1092-4388(2009/08-0243)

Hirata, Y., Whitehurst, E., & Cullings, E. (2007). Training native English speakers to identify Japanese vowel length contrast with sentences at varied speaking rates. *The Journal of the Acoustical Society of America*, *121*(6), 3837–3845. https://doi.org/10.1121/1.2734401

Huensch, A., & Tremblay, A. (2015). Effects of perceptual phonetic training on the perception and production of second language syllable structure. *Journal of Phonetics*, *52*, 105–120. https://doi.org/10.1016/j.wocn.2015.06.007

Iino, A. (2019). Effects of HVPT on perception and production of English fricatives by Japanese learners of English. In F. Meunier, J. Van de Vyver, L. Bradley & S. Thouësny (Eds.), *CALL and complexity – short papers from EUROCALL 2019* (pp. 186–192). Research-publishing.net.

Isbell, D. R. (2021). Can the test support student learning? Validating the use of a second language pronunciation diagnostic. *Language Assessment Quarterly*, *18*(4), 331–356. https://doi.org/10.1080/15434303.2021.1874382

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults. *The Journal of the Acoustical Society of America*, *118*(5), 3267–3278. https://doi.org/10.1121/1.2062307

Iverson, P., Pinet, M., & Evans, B. G. (2012). Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics*, *33*(1), 145–160. https://doi.org/10.1017/S0142716411000300

Jamieson, D. G., & Morosan, D. E. (1986). Training non-native speech contrasts in adults: Acquisition of the English /ð/-/θ/ contrast by francophones. *Perception & Psychophysics*, *40*(4), 205–215. https://doi.org/10.3758/BF03211500

Lambacher, S. G., Martens, W. L., Kakehi, K., Marasinghe, C. A., & Molholt, G. (2005). The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics*, *26*(2), 227–247. https://doi.org/10.1017/S0142716405050150

Lee, A. H., & Lyster, R. (2016). Effects of different types of corrective feedback on receptive skills in a second language: A speech perception training study. *Language Learning*, *66*(4), 809–833. https://doi.org/10.1111/lang.12167

Lee, H.-Y., & Hwang, H. (2016). Gradient of learnability in teaching English pronunciation to Korean learners. *The Journal of the Acoustical Society of America*, *139*(4), 1859–1872. https://doi.org/10.1121/1.4945716

Lengeris, A., & Nicolaidis, K. (2014). Effect of phonetic training on the perception of English consonants by Greek speakers in quiet and noise. *Proceedings of Meetings on Acoustics*, *22*(1), 1–6. https://doi.org/10.1121/2.0000025

Leong, C. X. R., Price, J. M., Pitchford, N. J., & Heuven, W. J. B. van. (2018). High variability phonetic training in adaptive adverse conditions is rapid, effective, and sustained. *PLOS ONE, 13*(10), e0204888. https://doi.org/10.1371/journal.pone.0204888

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/10.1121/1.408177

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y., Yamada, T. (1994). Training Japanese listeners to identify /r/ and /l/. III. Long-term retention of new phonetic categories. *Journal of the Acoustical Society of America*, *96*(4), 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. *The Journal of the Acoustical Society of America*, *89*(2), 874–886. https://doi.org/10.1121/1.1894649

Logan, J., & Pruitt, J. (1995). Methodological issues in training listeners to perceive non-native phonemes. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross language research* (pp. 351–378). York Press.

Melnik, G. A., & Peperkamp, S. (2021). High-Variability Phonetic Training enhances second language lexical processing: Evidence from online training of French learners of English. *Bilingualism: Language and Cognition*, *24*(3), 497–506. https://doi.org/10.1017/S1366728920000644

Mora, J. C., & Levkina, M. (2017). Task-based pronunciation teaching and research: Key issues and future directions. *Studies in Second Language Acquisition*, *39*(2), 381–399. https://doi.org/10.1017/S0272263117000183

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

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, *11*(2), 364–386. https://doi.org/10.1177/10944281062910

Nishi, K., & Kewley-Port, D. (2007). Training Japanese listeners to perceive American English vowels: Influence of training sets. *Journal of Speech, Language, and Hearing Research*, *50*(6), 1496–1509. https://doi.org/10.1044/1092-4388(2007/103)

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *British Medical Journal*. https://doi.org/10.1136/bmj.n71

Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (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

Pierrehumbert, J. B. (2002). Word specific phonetics. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology VII* (pp. 101–140). Mouton de Gruyter.

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

Qian, M. (2018). *An adaptive computational system for automated, learner-customized segmental perception training in words and sentences: Design, implementation, assessment* [Doctoral dissertation, Iowa State University].

Saito, K., Hanzawa, K., Petrova, K., Kachlicka, M., Suzukida, Y., & Tierney, A. (2022). Incidental and multimodal high variability phonetic training: Potential, limits, and future directions. *Language Learning*, *72*(4), 1049–1091. https://doi.org/10.1111/lang.12503

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*(3), 652–708. https://doi.org/10.1111/lang.12345

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

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics*, *3*(3), 243–261. https://doi.org/10.1017/S0142716400001417

Shinohara, Y., & Iverson, P. (2018). High variability identification and discrimination training for Japanese speakers learning English /r/–/l/. *Journal of Phonetics*, *66*, 242–251. https://doi.org/10.1016/j.wocn.2017.11.002

Shinohara, Y., & Iverson, P. (2021). The effect of age on English /r/-/l/ perceptual training outcomes for Japanese speakers. *Journal of Phonetics*, *89*, 101108. https://doi.org/10.1016/j.wocn.2021.101108

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*(2), 209–235. https://doi.org/10.1075/jslp.19010.sil

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning*, *60*(2), 263–308. https://doi.org/10.1111/j.1467-9922.2010.00562.x

Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception & Psychophysics*, *36*(2), 131–145. https://doi.org/10.3758/BF03202673

Suzukida, Y., & Saito, K. (2021). Which segmental features matter for successful L2 comprehensibility? Revisiting and generalizing the pedagogical value of the functional load principle. *Language Teaching Research*, *25*(3), 431–450. https://doi.org/10.1177/1362168819858

Thomson, R. I. (2011). Computer assisted pronunciation training: Targeting second language vowel perception improves pronunciation. *CALICO Journal*, *28*(3), 744–765. https://doi.org/10.11139/cj.28.3.744-765

Thomson, R. I. (2012a). English Accent Coach: Not quite a fairy godmother for pronunciation instruction, but a step in the right direction. *CONTACT*, *38*(1), 18–24.

Thomson, R. I. (2012b). Improving L2 listeners' perception of English vowels: A computer-mediated approach. *Language Learning*, *62*(4), 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*(2), 208–231. https://doi.org/10.1075/jslp.17038.tho

Thomson, R. I. (2022). Perception in pronunciation training. In J. Levis, T. M. Derwing & S. Sonsaat Hegelheimer (Eds.), *Pronunciation in second language learning and teaching: Innovations and developments in research and teaching* (pp. 42–60). Wiley.

Thomson, R. I. (2024). English Accent Coach Asteroids (1.0) [Mobile application software]. https://apps.apple.com/ca/app/english-accent-coach-asteroids/id1607018351

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*. Iowa State University.

Tyler, M. D. (2019). PAM-L2 and phonological category acquisition in the foreign language classroom. In A. M. Nyvad, M. Hejná, A. Højen, A. Jespersen, & M. H. Sørensen (Eds.), *A sound approach to language matters—In honor of Ocke-Schwen Bohn* (pp. 607–630). Aarhus University.

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

Uchihara, T., Webb, S., Saito, K., & Trofimovich, P. (2022). The effects of talker variability and frequency of exposure on the acquisition of spoken word knowledge. *Studies in Second Language Acquisition*, *44*(2), 357–380. https://doi.org/10.1017/S0272263121000218

Ueda, R., & Hashimoto, K. (2019). Perceptual training in a classroom setting: Phonemic category formation by Japanese EFL learners. In J. Levis, C. Nagle, & E. Todey (Eds.), *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference* (pp. 237–249). Iowa State University.

Vandergrift, L., & Baker, S. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning*, *65*(2), 390–416. https://doi.org/10.1111/lang.12105

Wang, X., & Munro, M. J. (2004). Computer-based training for learning English vowel contrasts. *System*, *32*(4), 539–552. https://doi.org/10.1016/j.system.2004.09.011

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*(6), 3649–3658. https://doi.org/10.1121/1.428217

Wiener, S., Chan, M. K., & Ito, K. (2020). Do explicit instruction and high variability phonetic training improve nonnative speakers' Mandarin tone productions? *The Modern Language Journal*, *104*(1), 152–168. https://doi.org/10.1111/modl.12619

Wong, J. W. S. (2014). The effects of high and low variability phonetic training on the perception and production of English vowels /e/-/æ/ by Cantonese ESL learners with high and low L2 proficiency levels. In H. Li & P. Ching (Eds.), *Proceedings of the 15th Annual Conference of the International Speech Communication Association* (pp. 524–528). International Speech Communication Association.

Yang, R., Nanjo, H., & Dantsuji, M. (2021). Self adaptive phonetic training for Mandarin nasal codas. *Computer-Assisted Language Learning Electronic Journal*, *22*(1), 391–413.

Yeon, S.-H. (2004). *Teaching English word-final alveolopalatals to native speakers of Korean* [Doctoral dissertation, University of Florida].

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

Zhang, X., Cheng, B., & Zhang, Y. (2022). A hands-on tutorial for systematic review and meta-analysis with example data set and codes. *Journal of Speech, Language, and Hearing Research*, *65*(9), 3217–3238. https://doi.org/10.1044/2022_JSLHR-21-0060

Cite this article: Uchihara, T., Karas, M., & Thomson, R. I. (2025). High variability phonetic training (HVPT): A meta-analysis of L2 perceptual training studies. *Studies in Second Language Acquisition*, 1–34. https://doi.org/10.1017/S0272263125100879