Korean Journal of English Language and Linguistics, **Vol 25, October 2025, pp. 1416-1443**
DOI: 10.15738/kjell.25..202510.1416

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474

http://journal.kasell.or.kr

The Efficacy of High Variability Phonetic Training for L2 Speech Perception in EFL Contexts: A Meta-Analytic Approach

Seyeon Choe · Hyoyoung Park · Hyunkee Ahn (Seoul National University)

Received: August 22, 2025
Revised: October 16, 2025
Accepted: October 22, 2025

Choe, Seyeon (First author)
Ph.D. student, Department of English Language Education
Seoul National University
1 Gwanak-ro, Gwanak-gu
Seoul, Korea
Email: caomei214@snu.ac.kr

Park, Hyoyoung (Co-author)
Ph.D. candidate, Department of English Language Education
Seoul National University
Email: hpark412@snu.ac.kr

Ahn, Hyunkee (Corresponding author)
Professor, Department of English Language Education
Seoul National University
Email: ahnhk@snu.ac.kr

ABSTRACT

Choe, Seyeon, Hyoyoung Park and Hyunkee Ahn. 2025. The efficacy of high variability phonetic training for L2 speech perception in EFL contexts: A meta-analytic approach. *Korean Journal of English Language and Linguistics* 25, 1416-1443

This meta-analysis examined the effectiveness of High Variability Phonetic Training (HVPT) in improving English L2 speech perception among EFL learners and examines how learner-, training-, and testing-related factors moderate its efficacy. Synthesizing data from 15 studies, the overall immediate effect of HVPT was small to medium (g = 0.62, 95% CI [0.53; 0.71]). HVPT effects diminished over time, with smaller effect sizes at one- to two- month follow-ups, suggesting limited long-term retention. Learner-related moderators revealed stronger effects for adolescents (g = 0.76), and pre-intermediate learners (g = 0.99). Training-related factors associated with greater effectiveness included real-word training stimuli (g = 0.69), exposure to more than ten talkers (g = 0.79), and inclusion of visual support during L2 speech training (g = 0.75). Shorter overall training durations (≤5 weeks; g = 0.73), and training sessions longer than 30 minutes (g = 0.75) were also linked to larger effects. Identification training was generally more effective than discrimination training (g = 0.62 vs. 0.33), yet training that combined both tasks yielded the largest effect (g = 0.96). Testing task type (identification vs. discrimination) showed little impact, but testing stimulus type (word vs. nonword) made a clear difference (g = 0.68 vs. 0.36). Transfer effects were stronger for word-based than nonword-based training, particularly in long-term retention (g = 0.98). Overall, the findings suggest that HVPT can produce meaningful improvements in L2 speech perception among EFL learners, particularly under specific learner and training conditions.

KEYWORDS

EFL, high variability phonetic training, HVPT, meta-analysis, L2 speech perception training

1. Introduction

High variability phonetic training (HVPT) is generally recognized as a form of perceptual training—primarily targeting segmentals—that exposes second language (L2) learners to a wide range of auditory stimuli produced by multiple speakers across diverse phonetic contexts (Thomson 2018). Research has demonstrated the effectiveness of HVPT in enhancing pronunciation skills among L2 learners, especially for adults (Barriuso and Hayes-Harb 2018, Mahdi and Mohsen 2024, Thomson 2018, Uchihara et al. 2024). Although English L2 learners may experience substantially different learning conditions depending on whether they study in an English as a foreign language (EFL) or English as a second language (ESL) context (Nayar 1997)—such as the amount and quality of input or opportunities for interaction, this distinction is often overlooked in discussions of L2 pronunciation teaching and learning, which tend to focus predominantly on learners in ESL environments. For instance, key variables such as length of residence (LOR) (e.g., Derwing 2008) or the discussion of the window of maximal opportunity (WMO) hypothesis (Derwing and Munro 2015) primarily pertain to ESL learners, with limited consideration of EFL contexts. Moreover, in a recent meta-analysis of HVPT, Uchihara et al. (2024) reported that the effect size was larger for foreign language contexts (Hedges' g = 0.60) than for second language contexts (Hedges' g = 0.34), suggesting that HVPT tends to be more effective for EFL learners than for ESL learners in enhancing L2 production. In this regard, the effectiveness of HVPT in EFL contexts warrants dedicated and systematic investigation.

HVPT is based on the theoretical premise that perception is a prerequisite for production (Flege 1995a, Thomson 2018, 2022a). Its core objective is to enhance learners' L2 *pronunciation* skills by first strengthening their *perceptual* abilities (Thomson 2018). In other words, the development of L2 pronunciation skills is contingent on the learner's capacity to accurately *perceive* novel or contrastive L2 speech sounds in the target language (Thomson 2022a). The role of perception in shaping production is well documented in the literature (see Kingston 2007). Houde and Jordan (1988, 2002) demonstrated through manipulated auditory feedback experiments that speakers adjust their articulation based on how their speech sounds, indicating an auditory-based control mechanism in speech production. Thomson (2022a) also emphasized the importance of L2 learners' *perceptual* reorientation in L2 pronunciation instruction, given that they already possess automatic L1-based speech perception systems.

Recognizing the foundational role of perception in guiding speech production, this study examines the effectiveness of HVPT in enhancing EFL learners' perceptual abilities and explores the key moderating factors influencing its outcomes, using a meta-analytic approach. In EFL contexts—where English is not an official language and opportunities for input and output are inherently limited (Nayar 1997)—a well-developed L2 sound perceptual system for L2 learners becomes particularly critical, as it can shape accurate pronunciation (e.g., Flege 1995a, Escudero 2005). While the ultimate goal of HVPT is to improve pronunciation, it is fundamentally a perceptual training method. Therefore, examining perceptual outcomes provides a valid and meaningful index of its impact. Accordingly, the present study focuses exclusively on *perceptual* outcomes, acknowledging their pivotal role in successful L2 pronunciation acquisition. In sum, this meta-analysis aims to quantitatively synthesize the effectiveness of HVPT on EFL learners' speech perception and to identify the key factors moderating this training effect.

2. Literature Review

2.1 Why Context Matters: Rethinking L2 Pronunciation Teaching in ESL and EFL Settings

While ESL and EFL contexts share the common goal of equipping learners to use English effectively outside the classroom, the distinction between them—though not strictly deterministic (Brown 2000, Nayar 1997)—entails important differences in instructional and learning practices that merit attention (Tanner and Henrichsen 2022). Previous meta-analyses of pronunciation instruction have reported separate effect sizes for EFL and ESL contexts, typically showing stronger effects in EFL settings (Choe et al. 2020, Lee et al. 2015, Uchihara et al. 2024). However, to our knowledge, these contextual differences have rarely been examined in depth. That is, contextual distinctions have often been conflated with other moderating variables, making it difficult to isolate the unique role of context when interpreting moderator effects. Nevertheless, in the context of L2 pronunciation pedagogy, two critical distinctions can be drawn between EFL and ESL environments. The first concerns learners' linguistic backgrounds: EFL learners typically share a common L1, resulting in similar pronunciation challenges, whereas ESL classrooms involve greater linguistic diversity and consequently more varied pronunciation difficulties (Tanner and Henrichsen 2022). The second pertains to communicative demands, which differ substantially between the two contexts (Nayar 1997). Specifically, while ESL learners are generally immersed in English-speaking environment, EFL learners have limited exposure to naturalistic input and often lack opportunities to test the intelligibility of their speech in authentic communicative situations.

These differences between ESL and EFL contexts align with a broader paradigm shift in L2 pronunciation research and pedagogy, particularly in how pronunciation goals are conceptualized and implemented across settings. Two key principles have shaped this shift: the Nativeness Principle and the Intelligibility Principle (Levis 2005, 2020). The former equates native-like pronunciation with intelligible speech, regardless of actual comprehensibility. In contrast, the now widely supported Intelligibility Principle prioritizes mutual understanding as the goal of communication (Munro and Derwing 1995, 2020). Empirical evidence indicates that even heavily accented speech can be both intelligible and comprehensible (Munro and Derwing 1995), prompting a move toward listener-based outcomes—such as intelligibility and comprehensibility—in L2 pronunciation teaching and learning (Levis 2020, Munro and Derwing 2015). This paradigm shift, however, may vary in its applicability across instructional contexts, given the persistent reliance on traditional, teacher-centered practices and learners' reluctance to engage in face-threatening interaction in many EFL contexts (Saito and Ebsworth 2004). From the learner's perspective, the need for the Intelligibility Principle may be less salient in EFL contexts, where English seldom functions as a medium of everyday communication. Furthermore, in settings where learners and teachers share the same L1, their mutual familiarity with L1-accented English may hinder valid and reliable assessment of L2 pronunciation progress (Foote et al. 2016). In this light, the EFL-ESL distinction should not be overlooked in L2 pronunciation pedagogy, particularly at the classroom level, where pedagogical conditions and learner expectations often diverge markedly (e.g., Saito and Ebsworth 2004).

As Levis (2020) emphasizes, intelligibility is inherently context dependent. This suggests that L2 pronunciation pedagogy should align with the communicative realities in which it is implemented. Building on this insight, the present study aims to explicitly distinguish EFL settings from ESL settings—not merely as moderator variables, but as core determinants of the pedagogical approach—in an effort to promote a more context-sensitive understanding of L2 pronunciation instruction. In this study, the EFL context is defined as "the role of English in countries where it is taught as a subject in schools but where it has no recognized status or function" (Nayar 1997, p.13), to ensure conceptual clarity in distinguishing it from the ESL context.

2.2 The Role of Perception in Shaping L2 Speech Production

Although perception and production are generally regarded as distinct yet interrelated processes, the question of which precedes the other in L2 acquisition remains a subject of ongoing debate (Escudero 2007, Flege 1995a, Flege and Bohn 2021, Thomson 2022b). Theoretically, the Speech Learning Model (SLM; Flege 1995a) and its revised version, SLM-r (Flege and Bohn 2021), differ in their accounts of how this sequencing develops. Flege's (1995a) SLM posits "that without accurate perceptual *targets* to guide the sensorimotor learning of L2 sounds, production of the L2 sounds will be inaccurate" (p. 238). In contrast, the revised SLM-r introduces a *co-evolution* hypothesis, proposing that perception and production mutually influence one another throughout L2 speech learning, in response to mixed findings in empirical research (Flege and Bohn 2021). While the SLM-r accommodates evidence suggesting that production may sometimes precede perception, it does not replace the original assumption with a unidirectional production-to-perception account. Rather, it reframes the original model as a bidirectional, co-evolutionary framework linking perception and production.

A substantial body of empirical research has shown that successful L2 speech learning is grounded in perceptual development, which paves the way for subsequent improvements in production (Georgiou 2022, Rato 2013, Sebastián-Gallés and Baus 2005). However, as acknowledged within the co-evolutionary perspective of SLM-r, some studies have reported cases in which learners' production abilities may, at times, surpass their perceptual skills (Goto 1971, Tsukada et al. 2005). These findings, however, have been questioned by Escudero (2007), who identified several methodological limitations in studies suggesting that production can precede perception—such as the artificial task conditions or the effects of prior articulatory training—potentially resulting in overly self-monitored and unnatural L2 speech. Likewise, Thomson (2022b) argues that, much like in first language acquisition, perception precedes production in L2 learning, and that studies reporting the opposite primarily reflect the influence of explicit knowledge in production rather than refuting this fundamental sequence. Notably, evidence from HVPT, which has been found to improve L2 production through perceptual training, provides compelling empirical support for the view that perception precedes production in L2 acquisition (Barriuso and Hayes-Harb 2018, Thomson 2022a, Uchihara et al. 2024). In this respect, "HVPT is likely to manifest in perception first, before affecting production" (Thomson 2018, p.214).

Although improving L2 production skills is the ultimate goal of HVPT, the path from perceptual gains to productive accuracy is far from straightforward, largely because speech production requires complex articulatory motor control involving muscle coordination, timing, and sensorimotor integration (e.g., Levelt et al. 1999). That is, there exists a somewhat asymmetrical relationship between perception and production (Sebastián-Gallés and Baus 2005, Thomson 2022b, Uchihara et al. 2024, Zhang et al. 2021). Against this backdrop, the present study focuses exclusively on the perceptual outcomes of HVPT. This emphasis on perception is not intended to diminish the importance of production outcomes; rather, it foregrounds perception as a necessary precursor to production in L2 speech learning, aiming for a more nuanced understanding of how HVPT facilitates L2 speech learning in EFL contexts.

2.3 The Effectiveness of HVPT and its Moderating Factors in Enhancing L2 Speech Perception

When L2 learners are exposed to highly variable speech input, they tend to develop more robust L2 phonological systems (Levis 2016). In this regard, HVPT has gained increasing attention as an effective approach to L2 speech learning, capitalizing on natural variability in speech input to enhance L2 learners' perception and production of target sounds (Barriuso and Hayes-Harb 2018, Levis 2016, Thomson 2018). This section synthesizes previous findings on how factors such as age, proficiency, training methods, and testing methods influence HVPT outcomes.

While HVPT studies suggest that adults typically benefit from training, children often exhibit variation in effectiveness depending on their age. Giannakopoulou et al. (2017) reported that across 10 training sessions for English tense-lax vowel distinctions involving minimal-pair picture-selection HVPT tasks, adults consistently outperformed children aged approximately 7.8 to 9.8 years. While both age groups improved on discrimination tests following training, children showed greater gains in the low-variability training condition than in the high-variability one—contrary to the authors' expectations. This pattern may be partly explained by age variation within child groups, as other studies have shown that older children (approximately 8-12 years) tend to outperform younger children (around 6-8 years) in L2 speech learning (Brekelmans et al. 2024, Shinohara and Iverson 2021). These findings suggest that HVPT studies should account for developmental differences within child populations, rather than treating children as a homogeneous group.

Research on HVPT has shown that while it benefits learners across all proficiency levels, its effect may vary depending on learners' initial proficiency levels. Wong's (2014) examined how learners with differing levels of perception and production abilities responded to HVPT targeting the English /e-æ/ contrast, revealing both groups showed comparable improvement after training. This finding suggests that intensive HVPT may offset the initial proficiency-based differences by directing learners' attention to phonetic cues, thereby facilitating the acquisition of difficult non-native contrasts regardless of prior skill level. On the contrary, Lee and Hwang's (2016) findings indicate that HVPT may yield differential benefits depending on learners' initial proficiency levels. Specifically, lower-level learners showed more pronounced gains in identification accuracy—particularly for consonants—than their higher-level counterparts. Thus, their study suggests that HVPT is most effective when informed by diagnostic insights into learners' current skills, particularly when training involves difficult contrasts.

Recent HVPT studies suggest that nonwords may offer distinct advantages over words as training ones (Ortega et al. 2021, Thomson and Derwing 2016). Thomson and Derwing (2016) compared a Phonetic Group, exposed to a mix of words and nonwords (mostly nonwords), with a Real Word Group trained exclusively on real words, targeting ten English vowels. They found that Phonetic Group showed greater gains in an elicited imitation task, which they interpret as evidence that nonwords facilitate more accurate encoding of segmental detail by minimizing reliance on lexical memory. Ortega et al. (2021) extended this evidence with advanced Spanish-Catalan bilinguals by training separate groups on either nonwords or real words for English /æ-ʌ/ contrast. The results showed that only the nonword group showed significant improvements in vowel production, as measured by acoustic distance from native targets, and these gains generalized to novel real-word items. The word group, by contrast, failed to improve. Taken together, these findings on the effectiveness of nonword training support the view that nonword stimuli in HVPT direct learners' attention to fine phonetic detail without lexical interference, thereby fostering more robust L2 phonological abstraction that facilitates more accurate production.

A growing body of HVPT research has aimed to determine not only whether training leads to improvement, but also whether such gains generalize to unfamiliar talkers and contexts and persist over time. These two dimensions of learning are known as *generalization*—the extent to which training effects transfer beyond familiar input—and *retention*—the extent to which those effects are maintained over time, typically assessed using delayed posttests (Thomson 2018). However, findings on the generalization and retention effects of HVPT remain inconsistent. Zhang et al. (2021), in their meta-analysis on HVPT, reported a large generalization effect size (g = 0.72, 95% CI [0.15; 1.29]) and an even larger retention effect (g = 1.09, 95% CI [0.39; 1.78]); however, as the authors noted, the latter should be interpreted with caution, given that it is based on only two studies. In contrast, Uchihara et al. (2024) reported more limited effects for both generalization and retention by comparing delayed posttest results to both pretest results and posttest outcomes. These divergent results suggest that the effects of HVPT on

generalization and retention warrant further reexamination with consistent operational definitions and outcome domains.

HVPT studies have consistently regarded variability across multiple talkers as a central component of the training paradigm (Logan et al. 1991, Uchihara et al. 2025, Wong 2014), as it leverages the variability in speech from different talkers to enhance learners' ability to generalize and adapt to new speech sounds (Barriuso and Hayes-Harb 2018, Thomson 2018). However, there is a lack of concrete empirical data to determine how many are truly required. In a recent meta-analysis, Uchihara et al. (2024) reported that the number of talkers—ranging from 3 to 30, with most studies using between 3 and 6—did not significantly predict learning outcomes. This suggests that increasing or decreasing talker variability within that range has little impact on the transfer of HVPT effects to production. In contrast, Mahdi and Mohsen (2024) found that studies involving five to eight talkers yielded a larger effect size ($g = 0.77$) than those with only two to four talkers ($g = 0.55$). This suggests a potential threshold effect, whereby moderate talker variability (e.g., five to eight talkers) may yield greater learning gains than lower levels of variability.

Although both Forced Choice Identification (FCID) and AXB discrimination tasks are commonly employed to train and assess speech perception in HVPT, they differ in the cognitive processes underlying speech perception that they tap into (Flege 1995b, Thomson 2022b). Wayland (2007) found that performance on FCID tasks does not align well with results from AXB discrimination formats. It suggests that these tasks differ in terms of cognitive or perceptual demands. This distinction is further supported by Iverson et al. (2012), who examined how HVPT affected French learners of English with different levels of L2 experience. While both experienced and inexperienced learners improved significantly on vowel identification tasks, their gains in category discrimination were minimal, and, critically, the two measures were uncorrelated. In this respect, it has been suggested that identification tasks facilitate abstract phonological categorization, while discrimination tasks rely more directly on acoustic sensitivity and reflect relatively surface-level processing (Cebrian et al. 2024, Iverson et al. 2012, Shinohara and Iverson 2018, Wayland 2007). Given these differences, Thomson (2022b) argues that FCID tasks may better reflect real-world listening conditions, where listeners typically identify rather than explicitly compare speech sounds.

In light of these considerations, this study investigates the following two research questions by means of a meta-analytic approach:

(1) What is the overall immediate effect of HVPT on EFL learners' English speech perception, and to what extent are these effects retained at delayed posttests?
(2) Which specific learner-, training-, and testing-related variables most strongly moderate the effectiveness of HVPT in EFL contexts?

3. Method

3.1 Search Strategy and Data Sources

To examine the effects of HVPT on the perception of L2 learners in EFL contexts, relevant literature was systematically retrieved using the Scopus database and Google Scholar via the *Publish or Perish* software (Harzing 2007). Scopus was selected for its comprehensive indexing of peer-reviewed journal articles, while Google Scholar was used to broaden the scope to include gray literature such as unpublished doctoral dissertations. The primary

keywords for search were "high variability phonetic training." To minimize the risk of omitting relevant studies, synonymous terms were also employed, covering language context (e.g., "L2," "second language"), training type (e.g., "phonetic training," "auditory training," "perceptual learning"), and variability (e.g., "stimulus variability," "talker variability"). In total, 13 keyword combinations were used (see Appendix A for the full list). The search yielded 303 articles from Scopus and 3,600 from Google Scholar, resulting in 3,903 articles. After removing duplicate articles which appeared in both databases redundantly, 1,141 articles remained for screening.

In addition to database searches, 141 studies included in four recent meta-analyses—Zhang et al. (2021; $n = 14$), Mahdi and Mohsen (2024; $n = 17$), Uchihara et al. (2021; $n = 79$), and Uchihara et al. (2024; $n = 31$)—were manually cross-checked and added to the initial study pool to ensure comprehensive coverage of potentially relevant literature. Although Mahdi and Mohsen (2024) originally reported that 18 studies were included in their meta-analysis, two issues were identified. First, only 17 references were listed in their reference section. The missing study was traced and included by backtracking from Figure 3 in their article. Second, one of the listed studies was itself another meta-analysis and therefore excluded from the current screening. Consequently, a total of 17 studies from Mahdi and Mohsen (2024) were included in the screening process.

3.2 Study Selection Process Following PRISMA Guideline

Following the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statement (Page et al. 2021), the first and second authors each screened half of the 1,141 articles from *Publish or Perish* independently. During the initial screening phase, 565 records were excluded after reviewing titles and abstracts. These included studies on pathology, neuroscience, and sound perception without any intervention, as well as studies published before 1991, the year of the seminal HVPT study (Logan et al. 1991). Among the remaining 576 records, 19 could not be retrieved in full-text form. Subsequently, the remaining 557 reports for Publish or Perish were evaluated for eligibility. In addition, the 141 studies from four previous meta-analyses were screened. Among these, 47 duplicates—either internal to the meta-analyses or overlapping with the *Publish or Perish* results—were excluded. This left 94 additional studies to be screened. In total, 651 studies were included in the screening process. During this process, studies that met the following exclusion criteria were removed:

(1) studies that are not focusing on an EFL context;
(2) studies lacking sufficient descriptive statistics to calculate overall effect sizes (e.g., means, standard deviations, or the number of participants);
(3) studies employing a within-group design without a control group;
(4) studies that did not include any perceptual outcome measures (e.g., those focusing exclusively on production);
(5) book chapters without empirical data;
(6) additional duplicates overlooked during the initial screening process;
(7) theses (master's or undergraduate), poster presentations not included in peer reviewed proceedings; and
(8) review articles and other meta-analyses.

The authors cross-checked each other's screened sets to verify inclusion decisions and resolve any discrepancies. After screening, studies were included if they met all the following criteria:

(1) English as the target language in an EFL context;
(2) the study reported perception learning outcomes;
(3) means and standard deviations for both experimental and control groups were reported;
(4) the comparability of experimental and control groups before treatment was ensured through random assignment or statistical analysis; and
(5) perceptual training stimuli is recorded by multiple human speakers.

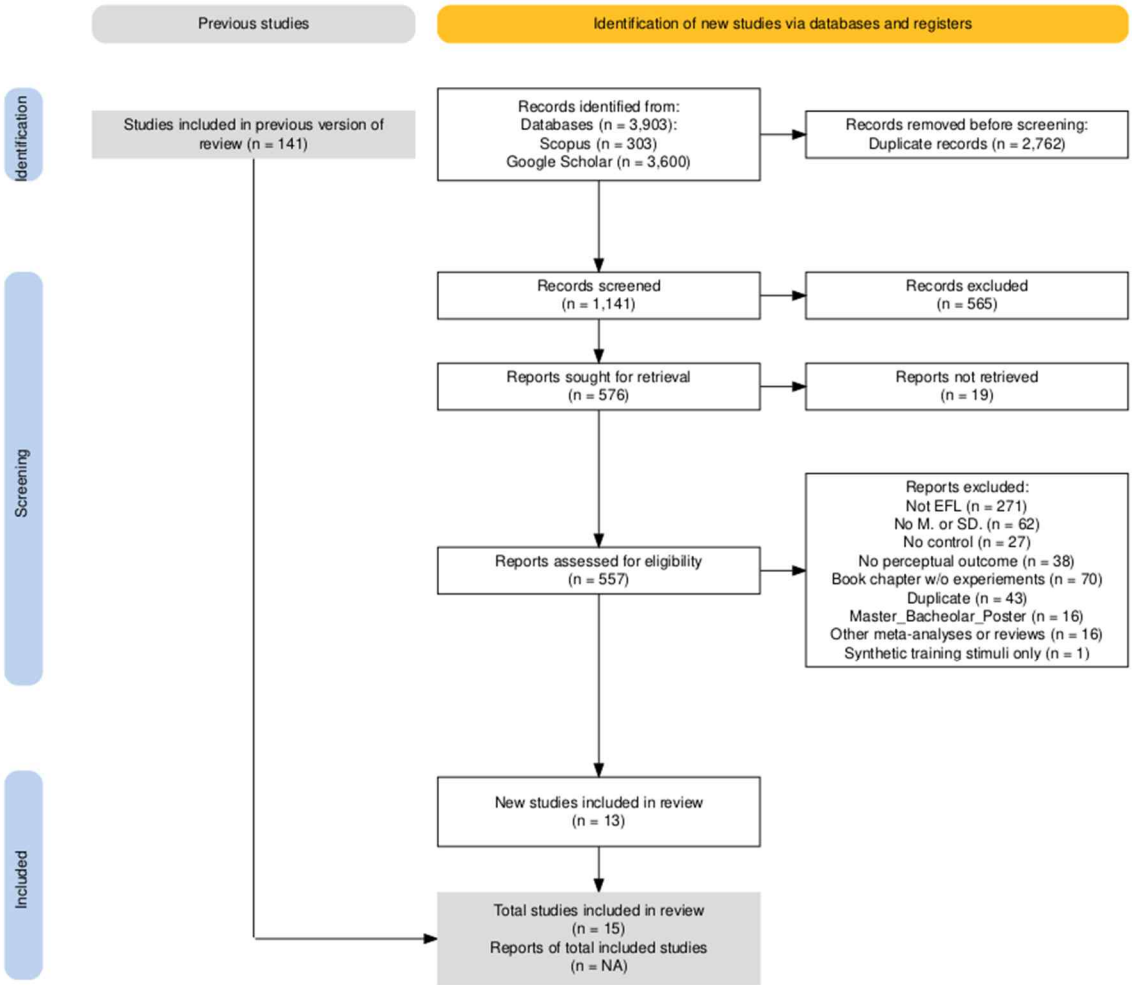


Figure 1. PRISMA Flowchart of the Study Selection Process

Consequently, a total of 15 studies met the exclusion and inclusion criteria and were retained for analysis. The 15 studies included in the analysis are presented in Appendix B. Of these, 13 were identified through *Publish or Perish* searches (N = 557). An additional two studies were retrieved from manual searches of four recent meta-analyses (N = 94) and are marked with a single asterisk (*). The studies retrieved through *Publish or Perish* were excluded for the following reasons: not conducted in an EFL context (n = 271), lack of necessary statistical parameters (n = 62), absence of a control group (n = 27), no perceptual outcome reported (n = 38), book chapters without experimental data (n = 70), duplicate studies that were missing during the initial duplicate screening (n = 43), theses that were not doctoral-level or poster presentations not included in peer-reviewed proceedings (n = 16),

meta-analyses or review papers (n = 16), and studies using only synthetic training stimuli (n = 1). The study selection process is illustrated in the PRISMA 2020 flowchart in Figure 1, generated using the *PRISMA2020 R package and Shiny App* (Haddaway et al. 2022). As the flowchart cannot fully illustrate the screening process of *Previous studies*, a brief explanation is provided here. Of the 94 studies retained after removing duplicates from four earlier meta-analyses, 52 were excluded for not focusing on an EFL context, 20 for missing statistical parameters, 14 for failing to include a control group, 5 for not reporting a perceptual outcome, and 1 for using no human-voice-based training. In total, 176 effect sizes (145 from immediate posttests and the remainder from delayed posttests) drawn from 15 independent studies were synthesized.

3.3 Data Extraction and Coding Procedures

To examine the factors that may moderate the effectiveness of HVPT on L2 speech perception, a range of potential moderator variables were identified and coded, as shown in Table 1. First, posttest results and delayed posttest results (after one month and after two months) were compared to assess the overall persistence of training effects over time. Subsequently, various moderators influencing posttest outcomes were examined further. These moderators were classified into three main categories: learner-related, training-related, and testing-related variables. For learner-related factors, we focused on age group (children, adolescents, or adults) and proficiency level (pre-intermediate, intermediate, upper-intermediate, or advanced). Training-related variables included the type of training task (identification, discrimination, or both), the type of stimuli used (word, nonword, or both), the number of talkers involved in the training (≤ 5, $5 < x \leq 9$, or $10 \leq$), the duration of the training program (in weeks) and length of each session (in minutes), the number of training sessions (≤ 5, $5 < x \leq 9$, or $10 \leq$), and training modality (perception only; perception and visual; perception and production; or perception, visual and production). For testing-related factors, we examined the task type (identification or discrimination), the word type (word or nonword), and the transfer, specifically whether the results reflected generalization or retention, based on performance with different types of training stimuli. In studies that reported separate outcomes for different target sound contrast, the data were coded for each sound pair or set. To differentiate these within-study outcomes, lowercase labels were used to identify these outcomes within the same study.

To facilitate consistent data coding, several secondary decisions were made during the coding process. First, as defined by the World Health Organization (n.d.), adolescence refers to the age range between 10 and 19 years in this study. Second, if information on training duration or length of each session was presented as a range rather than a specific value, the mean of the range was calculated and used in the coding. Third, studies that did not report values for a given moderator variable were left blank and excluded from the corresponding subgroup analysis. The selected studies were evenly divided between the first and second authors, who independently coded their assigned subsets using a shared Excel file on Google Drive. After completing the initial coding, the first author verified all coded data across three separate rounds, spaced one to two weeks apart, following the initial coding phase to enhance consistency and minimize potential errors.

Table 1. Coding Scheme for Extracted Variables

Variables		Value
Study	Author(Year)	Lowercase labels were used to categorize distinct outcomes reported in the same study. (e.g., Rato(2014)a)
Measurement time		Posttest Delayed Posttest (1mth) Delayed Posttest (2mths)
Learner	Age	Children Adolescents Adults
	Proficiency level	Pre-Intermediate Intermediate High-Intermediate Advanced
Training	Task type	Identification Discrimination Both
	Word type	Word Nonword Both
	The number of stimulus talkers	≤ 5 $5 < x \leq 9$ $10 \leq$
	Total duration (weeks)	≤ 5 $5 <$
	Duration per session (minutes)	≤ 30 $30 <$
	The number of training sessions	≤ 5 $5 < x \leq 9$ $10 \leq$
	Training modality	Perception only Perception and Visual Perception and Production Perception and Visual and Production
Testing	Task type	Identification Discrimination
	Word type	Word Nonword
	Transfer effects of training words	Word training_Generalization Word training_Retention Nonword training_Generalization Nonword training_Retention

3.4 Effect Size Calculation and Interpretation

Fifteen studies that included both control and experimental groups and reported means, standard deviations, and sample sizes were included in the analysis. The inclusion of a control group in intervention studies is essential for establishing whether the observed effects can be attributed to the instructional treatment (Thomson and Derwing 2015). Moreover, since perceptual outcomes were assessed using continuous measures on varying scales across studies, the standardized mean difference (SMD) was computed using Hedges' *g* to ensure comparability across studies (Beheshti et al. 2020). Hedges' *g* is a widely used SMD metric that adjusts for small sample bias, providing a more accurate estimate than Cohen's *d* (Cohen 1988), especially in studies with limited participant numbers (Borenstein et al. 2009, Hedges and Olkin 1985). The interpretation of effect sizes followed the L2 field-specific thresholds suggested by Plonsky and Oswald (2014). They suggested the magnitude of effect sizes around 0.4 is thought small, 0.7 medium, and 1.0 large for between group contrasts (Plonsky and Oswald 2014). All statistical analyses were carried out using *Meta-Mar v4.0.2* (Beheshti et al. 2020), a free AI-integrated web-based platform for meta-analysis. Effect sizes were synthesized with the *metacont* model, which is designed for continuous outcome data (Beheshti et al. 2020). Between-study heterogeneity was estimated with the restricted maximum-likelihood (REML) method (Viechtbauer 2005) and the Hartung-Knapp adjustment was applied to calculate robust confidence intervals for the random-effects model (Jackson et al. 2017, Sánchez-Meca and Marín-Martínez 2008). Prediction intervals were derived using a *t*-distribution. The prediction intervals were reported to complement confidence intervals as "[t]he confidence interval quantifies *the accuracy of the mean* while the prediction interval addresses the actual *dispersion of effect sizes*, and the two measure are not interchangeable" (Borenstein et al. 2009, p.131). Finally, all effect sizes were pooled using inverse-variance weighting to ensure appropriate weighting across studies.

4. Results

4.1 Publication Bias Analysis

A funnel plot was constructed for the posttest only to examine potential asymmetry in the distribution of effect sizes, as represented in Figure 2. Due to the small number of studies included in the delayed posttests (two or three studies), they were excluded from the publication bias analysis (Harrer et al. 2021). The funnel plot was generated using the *meta* package (Schwarzer et al. 2024) in *R* (R Core Team 2024). Publication bias was also assessed using multiple statistical methods, in line with the recommendation by Harrer et al. (2021) to combine approaches for more robust inference. It includes Egger's test, Thompson-Sharp test, trim-and-fill analysis, Begg's test, and three variants of the fail-safe N method.

Visual inspection of the contour-enhanced funnel plot showed slight asymmetry, particularly in the lower-right region where smaller studies with relatively large effect sizes clustered within the statistically significant range ($p < .01$). This pattern may indicate potential publication bias, as there was a noticeable absence of studies in the corresponding lower-left region that would reflect non-significant or negative results (Harrer et al. 2021). Complementing this visual assessment, Egger's regression test, however, indicated no evidence of funnel plot asymmetry ($t(143) = 0.04$, $p = .97$), and the Thompson-Sharp test likewise produced a non-significant result ($t(143) = -0.11$, $p = .92$), which suggests that there are no small-study effects based on regression diagnostics. Conversely, Begg's rank correlation test reached statistical significance ($z = 1.98$, $p = .048$). The trim-and-fill method also

estimated 13 potentially missing studies, and the adjusted mean effect size decreased from 0.62 to 0.53 (Δ = –0.07, ≈ –14% relative to the original), suggesting a modest publication bias favoring larger effects.

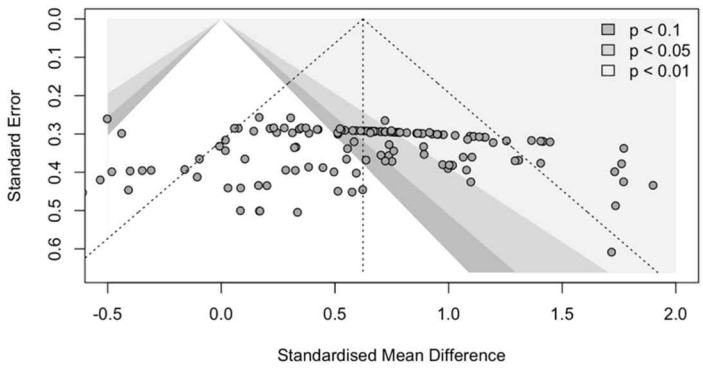


Figure 2. Contour-Enhanced Funnel Plot of Effect Sizes (posttest only)

Note. The contour-enhanced funnel plot illustrates that most studies cluster symmetrically around the overall mean, suggesting that the results are unlikely to be substantially affected by publication bias, despite minor asymmetry among smaller studies. The shaded areas correspond to levels of statistical significance (p < .1, .05, .01).

To further assess the robustness of the observed findings, three fail-safe N methods were conducted. Rosenthal's method estimated that 8,822 unpublished null-effect studies would be needed to render the overall result non-significant, which far exceeds the robustness threshold of 735 (i.e., $5k + 10$, where $k = 145$). According to Orwin's fail-safe N method, 145 additional null-effect studies would be needed to reduce the mean effect size from 0.62 to a trivial level of 0.31—still a sizable number given the assumption of zero effect. Rosenberg's fail-safe N, which incorporates both statistical significance and effect magnitude, yielded an estimate of 6,217, further reinforcing the conclusion that the meta-analytic result of this study is stable.

Taken together, while the results from Begg's rank correlation test and the trim-and-fill method suggest the possibility of publication bias, the high fail-safe N estimates across all three approaches support the robustness of the observed effect, even under conservative assumptions.

4.2 Meta-Analytic Results: Overall and Moderator Effects

The present meta-analysis synthesized data from 15 studies investigating the effectiveness of HVPT on L2 speech perception in EFL contexts. Of the 15 studies, four were doctoral dissertations, three were book chapters, seven were journal articles, and one was a poster presentation published in peer-reviewed conference proceeding. The calculated effect sizes are summarized in Table 2. In this study, only the results from the random-effects model are reported as confidence intervals derived from fixed-effects models are often too narrow when heterogeneity exists ($\tau^2 = 0.19$ [0.15; 0.30], $I^2 = 61.3\%$ [54.5%; 67.1%]), leading to undercoverage of the true effect size (Sánchez-Meca and Marín-Martínez 2008). Forest plots illustrating the results by study outcomes are provided in Appendix C (generated with *Meta-Analysis Online*; Fekete and Gyorffy 2025, https://metaanalysisonline.com/), and forest plots of HVPT effects by moderating variables are presented in Appendix D (produced in *R*; R Core Team 2024). Four moderator sublevels—*children* (age group; Shum et al. 2021a/b/c/d), *intermediate* (proficiency level; Li 2015a/b), *both* (training stimuli word type; Li 2015a/b), and *perception and production* (training modality;

Lacabex and Gallardo del Puerto 2014c/d)—were excluded from the analysis as each was represented by a single study.

Effect sizes were examined across three measurement time points: immediate posttest, 1-month delayed posttest, and 2-month delayed posttest. The pooled effect size was small (g = 0.59, 95% CI [0.51; 0.68]). For the posttest only, the effect size was slightly higher than the overall effect size, but still small-to-medium (g = 0.62, 95% CI [0.53; 0.71]) with moderate heterogeneity (τ^2 = 0.18). For the one-month delayed posttest, based on two studies (k = 2), the effect size decreased to small magnitude (g = 0.44, 95% CI [0.23; 0.64]), with low heterogeneity (τ^2 < 0.001). The effect size at the two-month delayed posttest (k = 3) was small (g = 0.37, 95% CI [0.02; 0.72]) with substantial heterogeneity (τ^2 = 0.42). The results indicate that while training yields clear immediate benefits, the magnitude of these effects tends to decline slightly after a month. The prediction interval indicates that the true effect in future comparable studies may vary substantially as it includes the possibility of null or negative effects, [–0.28; 1.47].

All subsequent subgroup analyses pertaining to learner, training, and testing factors were conducted using posttest data only, with the exception of the transfer effect. First, regarding learner-related moderators, results by age group revealed a stronger training effect for adolescents (g = 0.76 (medium to large), 95% CI [0.38; 1.17], τ^2 = 0.15) than adults (g = 0.63 (small to medium), 95% CI [0.54; 0.72], τ^2 = 0.43). Although adolescents appear to benefit more than adults, the findings indicate that HVPT is generally effective for both age groups. The prediction interval for the age subgroup ranges from –0.20 to 1.47, which suggests that the training effect may vary across studies, with some showing little or no improvement. For proficiency level, effect sizes were the largest among pre-intermediate learners (g = 0.99 (large), 95% CI [0.49; 1.48], τ^2 < 0.001), although this estimate was based on a small number of studies (k = 2). This was followed by advanced learners (g = 0.75 (medium to large), 95% CI [0.68; 0.83], τ^2 = 0.42) and upper-intermediate learners (g = 0.60 (small to medium), 95% CI [0.41; 0.80], τ^2 = 0.19). The results indicate that learners at all proficiency levels can benefit from HVPT, with particularly strong effects observed among pre-intermediate learners. The prediction interval for the proficiency-level subgroup (0.18–1.28) indicates that in future studies, the true effect of HVPT for learners at similar proficiency levels is likely to be positive, since the interval stays entirely above zero.

Training-related moderators yielded the following patterns in effect size estimates. Regarding task type, identification-only tasks yielded a notably larger effect (g = 0.62, 95% CI [0.52; 0.71], τ^2 = 0.17) than discrimination-only tasks (g = 0.33, 95% CI [0.10; 0.55], τ^2 = 0.06). This clear difference underscores the stronger contribution of identification training in HVPT. Moreover, training that involved both identification and discrimination tasks showed the largest effect size (g = 0.96, 95% CI [0.65; 1.27], τ^2 = 0.26). Although the three combined-task training studies under analysis varied in total training duration (ranging from 3 to 12 hours), their total exposure time fell within the overall range of included studies. Therefore, the large effect size appears to reflect influences beyond mere differences in training duration. Overall, the results suggest that identification training tend to be more effective than discrimination training in HVPT, and that combining both tasks yields the greatest overall benefit. However, the prediction interval for the training task type subgroup ranged from –0.23 to 1.48, which includes zero, indicating the possibility of null or negative effects. With respect to word type, training using real words yielded a medium effect (g = 0.69, 95% CI [0.60; 0.77], τ^2 = 0.11), while nonword-based training resulted in a small-to-medium effect (g = 0.50, 95% CI [0.29; 0.70], τ^2 = 0.20). This indicates that word-based training stimuli may be more effective than nonword stimuli for EFL learners in HVPT. The prediction interval for the training word type subgroup ranged from –0.08 to 1.38. The inclusion of zero in the interval suggests that future studies could report outcomes anywhere from no observable effect to meaningful improvements.

Table 2. Summary of Effect Sizes (Random-Effects Model)

Variables		k_s	k_o	SMD (Hedges' g)	95% CI	p	Prediction Interval	τ^2	Q
MEASUREMENT TIME									
	OVERALL	15	176	0.59	[0.51; 0.68]	< 0.0001	[-0.28; 1.47]	0.19	4.13
	Posttest	15	145	0.62	[0.53; 0.71]			0.18	
	Delayed Posttest (1mth)	2	12	0.44	[0.23; 0.64]			<0.001	
	Delayed Posttest (2mths)	3	19	0.37	[0.02; 0.72]			0.42	
LEARNER									
Age	OVERALL	14	141	0.64	[0.55; 0.73]	< 0.0001	[-0.20; 1.47]	0.17	0.52
	Adolescents	4	15	0.76	[0.38; 1.17]			0.15	
	Adults	13	126	0.63	[0.54; 0.72]			0.43	
Proficiency	OVERALL	9	104	0.73	[0.65, 0.81]	< 0.0001	[0.18; 1.28]	0.07	2.93
	Pre-intermediate	2	8	0.99	[0.49, 1.48]			<0.001	
	Upper-intermediate	4	34	0.60	[0.41; 0.80]			0.19	
	Advanced	3	62	0.75	[0.68; 0.83]			0.42	
TRAINING									
Task type	OVERALL	15	145	0.62	[0.53; 0.71]	< 0.0001	[-0.23; 1.48]	0.23	11.04 **
	Identification	12	114	0.62	[0.52; 0.71]			0.17	
	Discrimination	4	15	0.33	[0.10; 0.55]			0.06	
	Both	3	16	0.96	[0.65; 1.27]			0.26	
Word type	OVERALL	14	143	0.65	[0.57; 0.73]	< 0.0001	[-0.08; 1.38]	0.13	2.88
	Word	10	111	0.69	[0.60; 0.77]			0.11	
	Nonword	4	32	0.50	[0.29; 0.70]			0.20	
Number of talkers	OVERALL	15	145	0.62	[0.53; 0.71]	< 0.0001	[-0.23; 1.48]	0.18	17.25 ***
	≤ 5	10	63	0.40	[0.23; 0.57]			0.34	
	5 < x ≤ 9	4	18	0.69	[0.42; 0.95]			0.17	
	10 ≤	2	64	0.79	[0.71; 0.86]			<0.001	
Total Duration (week)	OVERALL	12	123	0.68	[0.60; 0.76]	< 0.0001	[0.01; 1.35]	0.11	2.92
	≤ 5	6	90	0.73	[0.64; 0.81]			0.05	
	5 <	6	33	0.51	[0.28; 0.74]			0.33	

Duration per Session (minute)	OVERALL	14	140	0.63	[0.53; 0.72]	< 0.0001	[-0.25; 1.50]	0.19	9.49 **
	≤ 30	8	66	0.46	[0.30; 0.62]			0.30	
	30 <	6	74	0.75	[0.65, 0.85]			0.55	
Number of sessions	OVERALL	15	145	0.62	[0.53; 0.71]	< 0.0001	[-0.23; 1.48]	0.18	0.14
	≤ 5	6	40	0.64	[0.46; 0.83]			0.20	
	5 < x ≤ 9	7	27	0.58	[0.28; 0.88]			0.48	
	10 ≤	3	78	0.64	[0.54; 0.73]			0.10	
Training modality	OVERALL	15	143	0.62	[0.53; 0.71]	< 0.0001	[-0.22; 1.45]	0.18	8.58 **
	Perception only	11	75	0.49	[0.35; 0.63]			0.24	
	Perception and Visual	3	40	0.75	[0.58; 0.91]			0.18	
	Perception and Visual and Production	2	28	0.73	[0.62; 0.85]			0.01	
TESTING									
Task type	OVERALL	15	145	0.62	[0.53; 0.71]	< 0.0001	[-0.23; 1.48]	0.18	0.05
	Identification	13	109	0.62	[0.51; 0.73]			0.22	
	Discrimination	3	36	0.64	[0.49; 0.78]			0.10	
Word type	OVERALL	15	145	0.62	[0.47; 0.69]	< 0.0001	[-0.37; 1.53]	0.18	4.65 *
	Word	12	117	0.68	[0.59; 0.77]			0.11	
	Nonword	5	28	0.36	[0.06; 0.65]			0.42	
Transfer effect	OVERALL	10	131	0.60	[0.51; 0.70]	< 0.0001	[-0.22; 1.42]	0.17	28.62 ****
	Word_Generalization	5	74	0.72	[0.62; 0.83]			0.12	
	Word_Retention	2	11	0.98	[0.62; 1.33]			0.13	
	Nonword_Generalization	4	26	0.40	[0.19; 0.61]			0.16	
	Nonword_Retention	3	20	0.18	[-0.03; 0.39]			0.07	

Notes. k_s = number of studies; k_o = number of observed outcomes across studies; Significance codes for *p*-values of *Q*: * = $p < .05$; ** = $p < .01$; *** = $p < .001$; **** = $p < .0001$

Analyses by number of talkers revealed that training with more than 10 talkers resulted in the medium to large effect size (g = 0.79, 95% CI [0.71; 0.86], $\tau^2 < 0.001$), followed by 6-9 talkers (g = 0.69, 95% CI [0.42; 0.95], $\tau^2 = 0.17$) and less than 5 talkers (g = 0.40, 95% CI [0.23; 0.57], $\tau^2 = 0.34$). The results indicate that exposure to a greater number of speakers during training may lead to more effective L2 speech learning for EFL learners. The prediction interval for the number of talkers subgroup ranged from –0.23 to 1.48, suggesting considerable uncertainty, with potential outcomes ranging from negligible to substantial effects. In terms of total training duration (in weeks), studies with shorter durations (≤ 5 weeks) yielded a stronger effect (g = 0.73 (medium), 95% CI [0.64; 0.81], $\tau^2 = 0.05$) compared to those with longer durations (5 < weeks), which shows small to medium effect (g = 0.51, 95% CI [0.28; 0.74], $\tau^2 = 0.33$). The prediction interval for training duration ranged from 0.01 to 1.35, indicating that future studies are likely to observe positive effects. When training duration per session was examined, training sessions longer than 30 minutes resulted in a medium-to-large effect (g = 0.75, 95% CI [0.65; 0.85], $\tau^2 = 0.55$), while shorter sessions (≤ 30 minutes) yielded a small to medium effect (g = 0.46, 95% CI [0.30, 0.62], $\tau^2 = 0.28$). The results suggest that providing EFL learners with longer exposure during each training session may lead to better learning outcomes. The prediction interval [–0.25; 1.50] indicates a wide range of possible future study outcomes depending on session duration, including null effects.

When examining the number of training sessions, all three subgroups exhibited comparable small-to-medium effect sizes, with minor variations across session counts. The highest effects were observed in studies with fewer than five sessions (g = 0.64, 95% CI [0.46; 0.83], $\tau^2 = 0.20$) and those with ten sessions (g = 0.64, 95% CI [0.54, 0.73], $\tau^2 = 0.10$), whereas studies involving six to nine sessions showed a slightly smaller effect (g = 0.58, 95% CI [0.28; 0.88], $\tau^2 = 0.48$). The prediction interval for the number of sessions ranged from –0.23 to 1.48. Since the interval includes zero, future studies may yield mixed outcomes—some showing limited effects, while others demonstrate substantial improvement. This finding suggests that the number of training sessions alone provides limited insight into the effectiveness of HVPT. When examined in conjunction with session length and total training duration, however, a more comprehensive understanding of learning outcomes may emerge. With regard to training modality, perception only training showed the smaller effect size (g = 0.49 (small to medium), 95% CI [0.35; 0.63], $\tau^2 = 0.24$) than perception and visual training (g = 0.75 (medium to large), 95% CI [0.58; 0.91], $\tau^2 = 0.18$) or perception with visual and production training (g = 0.73 (medium), 95% CI [0.62; 0.85], $\tau^2 = 0.01$). The results indicate that perception-based tasks alone may not be sufficient for optimal L2 speech perception improvement. Rather, the integration of additional modalities, visual or production components, appears to yield greater benefits. The prediction interval (–0.22 to 1.45), which indicate that while training modality is generally effective (g = 0.62), future studies may yield a wide range of outcomes—including null effects in some cases.

Testing-related moderators yielded the following patterns in effect size estimates. Notably, both identification and discrimination tests yielded comparable small-to-medium effect sizes, with small to medium effect for identification (g = 0.62, 95% CI [0.51; 0.73], $\tau^2 = 0.22$) and for discrimination (g = 0.64, 95% CI [0.49; 0.78], $\tau^2 = 0.10$). The comparable observed effect sizes across identification and discrimination testing tasks indicate that the choice of test task type may have minimal influence on the overall effectiveness of HVPT. However, this apparent similarity should be interpreted with caution, as the current analysis did not directly examine potential cross-task transfer effects, and the observed equivalence may partly reflect offsetting task-specific influences. The prediction interval (–0.23 to 1.48) for test task type suggests that some future studies may yield non-significant results. In terms of test word type, real-word-based tests yielded a medium effect size (g = 0.68, 95% CI [0.59; 0.77], $\tau^2 = 0.11$), whereas nonword-based tests produced a smaller effect (g = 0.36, 95% CI [0.06, 0.65], $\tau^2 = 0.42$). The prediction interval for test word type [–0.37; 1.53] includes zero, indicating that future studies may yield widely varying outcomes, including null or even negative effects.

Lastly, the subgroup analysis of transfer type revealed that generalization following word-based training yielded a medium effect (g = 0.72, 95% CI [0.62; 0.83], τ^2 = 0.12), and retention following word training showed a large effect (g = 0.98, 95% CI [0.62; 1.33], τ^2 = 0.13). In contrast, generalization after nonword training produced a smaller effect (g = 0.40, 95% CI [0.19; 0.61], τ^2 = 0.16), and retention after nonword training yielded a very small and statistically non-significant effect (g = 0.18, 95% CI [–0.03; 0.39], τ^2 = 0.07). These findings suggest that, for EFL learners, real-word training stimuli are more effective than nonword stimuli in promoting generalization to unfamiliar talkers and novel lexical items as well as in supporting the long-term retention of training effects. The prediction interval [–0.22; 1.42] for training word type in transfer effects includes zero, suggesting that the observed transfer outcomes may be null in some future studies.

5. Discussion and Conclusion

This meta-analysis synthesized data from 15 studies investigating the effectiveness of HVPT in EFL contexts by addressing two questions: (1) the immediate and delayed effects of HVPT on English speech perception among EFL learners, and (2) the extent to which learner-related, training-related, and testing-related factors moderated its effectiveness. Overall, this study indicates that HVPT is moderately effective for enhancing L2 speech perception in EFL contexts, although its long-term impact appears diminished. This observed decline over time (retention effect) is consistent with Uchihara et al. (2024). The pattern observed in this study may reflect natural attrition in newly acquired perceptual skills when they are not continuously reinforced through practice and exposure. Moreover, the finding that the effect size remained relatively stable between one and two months after training suggests that HVPT outcomes, while showing an initial decline, may reach a point of relative stability thereafter, although additional evidence from longer-term follow-up tests is needed.

With respect to learner-related factors, pre-intermediate learners showed the largest mean effects among the proficiency levels examined, and adolescents (10-19 years) exhibited the greatest effects among the age groups. This pattern suggests that HVPT tends to be more effective for learners with lower proficiency and for younger participants. However, because this study did not include beginner-level learners or younger children, it remains uncertain whether the *younger-the-better* effect extends to HVPT. Indeed, some studies have reported that young children benefit more from low-variability rather than high-variability input (Brekelmans et al. 2024, Giannakopoulou et al. 2017, Shinohara and Iverson 2021). Moreover, HVPT produced approximately medium-sized effects for higher proficiency levels and adult learners, indicating that the benefits are not confined to specific age subgroups. Overall, these findings suggest that HVPT exerts positive effects across different age ranges (likely after approximately age 10) and proficiency levels, consistent with previous research (Giannakopoulou et al. 2017, Hardison 2003).

As shown in previous research, stronger learning effects were observed in HVPT programs that employed identification tasks rather than discrimination tasks alone (Cebrian et al. 2024, Iverson et al. 2012, Shinohara and Iverson 2018, Thomson 2018, 2022b). Notably, the present study found the greatest benefits when both task types were combined. This pattern suggests that identification and discrimination tasks engage distinct perceptual mechanisms (Wayland 2007, Iverson et al. 2012) and yield complementary benefits when integrated into training. With respect to testing-related factors, the type of test (identification vs. discrimination) did not significantly influence EFL learners' performance. Still, previous studies on cross-task effects have indicated that specific training tasks can differentially affect performance across different types of testing tasks (Cebrian et al. 2024, Iverson et al. 2012). Thus, the negligible effect of test-task type observed here may be due to the present study's focus on overall effect sizes across tasks rather than the specific match or mismatch between training and testing task types.

The results indicated that extended phonetic training sessions (35-120 minutes), delivered over a condensed 3- to 5-week period, were particularly effective in promoting L2 speech *perception* gains. This finding contrasts with Uchihara et al. (2024), who suggested that 20-30 minutes might be the optimal duration for promoting L2 *production* development. One simple explanation is that whereas Uchihara et al. (2024) examined production, we focused on perceptual outcomes. Another possible explanation for this discrepancy is that over half of the participants in this study were at the upper-intermediate level or higher. Their prior exposure likely stabilized L2 phonetic categories, which, in turn, allowed high-density input to consolidate perceptual phonetic mappings over a short period, thereby addressing residual L2 speech learning difficulties among advanced learners (Pennington 2021). While short-term, intensive HVPT can produce moderate immediate L2 perception gains, evidence from our first research question also indicates that these gains may diminish over time. In this respect, identifying the effective booster-session schedule to sustain long-term L2 speech perceptual learning is crucial but remains under-specified. Moreover, it is unclear whether such scheduling interacts with learners' linguistic proficiency (Pennington 2021), a question that warrants further systematic investigation.

Combining multiple phonetic training modalities—such as *Perception and Visual* or *Perception and Visual and Production*—was associated with greater improvement. This finding underscores the importance of multimodal input in HVPT and is consistent with previous research (Hardison 2003, 2005, Hardison and Pennington 2021). Within this multimodal framework, the present results underscore the distinct contribution of visual (facial) cues: the *Perception and Visual* condition yielded effect sizes comparable to those of *Perception and Visual and Production*, suggesting no additional perceptual advantage from including the production component in HVPT. Overall, this pattern aligns with broader evidence that visible articulatory cues, such as lip gestures, facilitate perception and can also support production through transfer (Hardison 2003, Hardison and Pennington 2021). As Hardison and Pennington (2021, p. 70) note, "[w]ork on visible articulatory gestures can, therefore, be recommended as a way to improve learners' auditory perception of L2 sound contrast."

Although the present study found that training effectiveness increased with the number of talkers included in the input (ranging from 2 to 12), this contrasts with the recent meta-analysis by Uchihara et al. (2025), which reported no statistically significant linear relationship between talker number and perceptual gain across a broader range (2 to 30). This discrepancy can be interpreted in two ways. First, the maximum number of talkers examined here was only twelve, and Uchihara et al. (2025) noted that effect sizes tend to decline when more than ten talkers are included. Second, given that this study focuses on EFL contexts—where opportunities for authentic interaction and overall exposure to the target language are limited—increased talker variability may serve as a compensatory mechanism for such input scarcity.

The present study indicates that, within HVPT, word-based training stimuli were associated with larger perceptual gains than nonword stimuli—both immediately and in terms of generalization to new talkers or contexts and longer-term retention. This pattern contrasts with findings reported by Thomson and Derwing (2016) and Ortega et al. (2021). The most plausible explanation is that these earlier studies assessed production only, whereas the present study focused exclusively on perceptual outcomes. Moreover, considering that Ortega et al.'s (2021) work was conducted with EFL learners, the findings from this study indicate that perceptual ability—unlike production ability—can be more effectively enhanced through word-based stimuli in HVPT. In other words, perception and production may be partially dissociable in how they benefit from lexical context: lexical scaffolding may support perceptual encoding (Mora 2005, Rato and Carlet 2020), whereas nonword training materials may better isolate articulatory targets in production (Ortega et al. 2021).

The present study has several limitations that should be acknowledged. First, nearly one-third of the parameters included in the analysis (30.68%, 54 out of 176) were derived from a single doctoral dissertation (Aliaga-García 2017). This study reported results of 11 individual vowels (/iː, ɪ, e, ɜː, æ, ʌ, ɑː, ɒ, ɔː, ʊ, uː/)

and employed both identification (ID) and articulation (ART) training across identification and discrimination test types, which yielded a relatively large number of parameter estimates. Future meta-analyses should reconsider treating each vowel's results separately and instead explore alternative approaches, such as grouping sounds into broader phonetic categories (e.g., front vowels, back vowels).

Second, although HVPT is ultimately intended to enhance L2 speech production, production outcomes were not examined in the present study. This decision was primarily motivated by the unresolved nature of the mechanisms linking perceptual training to production improvement (Levelt et al. 1999, Sebastián-Gallés and Baus 2005) and by the assumption that it is more appropriate to interpret the effects of perceptual training within the perceptual domain. Nevertheless, incorporating production measures in future meta-analysis could provide greater opportunities for an in-depth exploration of the relationship between perception and production, which is closely related to the theoretical underpinnings of HVPT.

Despite these limitations, the present study demonstrates that HVPT is effective in enhancing EFL learners' L2 speech perception and identifies key moderating variables with implications for L2 speech learning in EFL environments.

References

Aliaga-García, 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.

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

Beheshti, A., M. L. Chavanon and H. Christiansen. 2020. Emotion dysregulation in adults with attention deficit hyperactivity disorder: a meta-analysis. *BMC Psychiatry* 20, 1-11. [Online platform]. Available online at https://www.meta-mar.com

Borenstein, M., L. V. Hedges, J. P. T. Higgins and H. R. Rothstein. 2009. *Introduction to Meta-Analysis*. Chichester, Wiley.

Brekelmans, G., B. G. Evans and E. Wonnacott. 2024. Training child learners on nonnative vowel contrasts with phonetic training: the role of task and variability. *Language Learning*. 1-36.

Brown, H. D. 2000. *Teaching by Principles: An Interactive Approach to Language Pedagogy* (2nd ed.). Longman.

Cebrian, J., N. Gavaldà, C. Gorba and A. Carlet. 2024. Differential effects of identification and discrimination training tasks on L2 vowel identification and discrimination. *Studies in Second Language Acquisition* 46(4), 1069-1093.

Choe, S., K. Lee and Y. So. 2020. The effects of phonemic awareness instructions on L2 listening comprehension: A meta-analysis. *Journal of Asia TEFL* 17(4), 1294-1309.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum.

Derwing, T. M. and M. J. Munro. 2015. *Pronunciation Fundamentals: Evidence-Based Perspectives for L2 Teaching and Research*. John Benjamins.

Derwing, T. M., M. J. Munro and R.I. Thomson. 2008. A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics* 29(3), 359-380.

Escudero, P. 2005. *Linguistic Perception and Second Language Acquisition: Explaining the Attainment of Optimal Phonological Categorization.* Doctoral dissertation, Utrecht University.

Escudero, P. 2007. Second-language phonology: The role of perception. In M. Pennington., ed., *Phonology in Context*, 109-134. Palgrave Macmillan.

Fekete, J. T. and B. Gyorffy. 2025. *MetaAnalysisOnline.com: An online tool for the rapid meta-analysis of*

clinical and epidemiological studies. [Online platform]. Available online at https://metaanalysisonline.com/

Flege, J. E. 1995a. Second language speech learning: Theory, findings, and problems. *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* 92(1), 233-277.

Flege, J. E. 1995b. Two procedures for training a novel second language phonetic contrast. *Applied Psycholinguistics* 16, 425-442.

Flege, J. E. and O. S. Bohn. 2021. The revised Speech Learning Model (SLM-r): Learning L2 sounds in a new language. In R. Wayland., ed., *Second Language Speech Learning: Theoretical and Empirical Progress*, 84-118. Cambridge University Press.

Foote, J. A., P. Trofimovich, L. Collins and F. S. Urzúa. 2016. Pronunciation teaching practices in communicative second language classes. *The Language Learning Journal* 44(2), 181-196.

Georgiou, G. P. 2022. The impact of auditory perceptual training on the perception and production of English vowels by Cypriot Greek children and adults. *Language Learning and Development* 18(4), 379-392.

Giannakopoulou, A., H. Brown, M. Clayards and E. Wonnacott. 2017. High or low? comparing high and low-variability phonetic training in adult and child second language learners. *PeerJ* 5. 1-35.

Goto, H. 1971. Auditory perception by normal Japanese adults of the sounds “L” and “R”. *Neuropsychologia* 9, 317-323.

Haddaway, N. R., M. J. Page, C. C. Pritchard and L. A. McGuinness. 2022. PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. *Campbell Systematic Reviews* 18, e1230.

Harrer, M., P. Cuijpers, T. A. Furukawa and D. D. Ebert. 2021. *Doing Meta-Analysis with R: A Hands-On Guide*. Chapman and Hall/CRC Press.

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

Hardison, D. M. 2005. Second-language spoken word identification: Effects of perceptual training, visual cues, and phonetic environment. *Applied Psycholinguistics* 26(4), 579-596.

Hardison, D. M. and M. C. Pennington. 2021. Multimodal second-language communication: Research findings and pedagogical implications. *RELC Journal* 52(1), 62-76.

Harzing, A. W. 2007. *Publish or perish*. [Computer software]. Available online at https://harzing.com/resources/publish-or-perish

Hedges, L. V. and I. Olkin. 1985. *Statistical Methods for Meta-Analysis*. Academic Press.

Houde, J. F. and M. I. Jordan. 1998. Sensorimotor adaptation in speech production. *Science* 279(5354), 1213-1216.

Houde, J. F. and M. I. Jordan. 2002. Sensorimotor adaptation of speech I: Compensation and adaptation. *Journal of Speech, Language, and Hearing Research* 45, 295-310.

Iverson, P., M. Pinet and B. G. Evans. 2012. Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics* 33, 145-160.

Jackson, D., M. Law, G. Rücker and G. Schwarzer. 2017. The Hartung-Knapp modification for random-effects meta-analysis: A useful refinement but are there any residual concerns?. *Statistics in Medicine* 36(25), 3923-3934.

Kingston, J. 2007. The phonetics-phonology interface. In P. de Lacy, ed., *The Cambridge Handbook of Phonology,* 401-434. Cambridge University Press.

Lacabex, E. G. and F. Gallardo del Puerto. 2014. Two phonetic-training procedures for young learners: Investigating instructional effects on perceptual awareness. *The Canadian Modern Language Review* 70(4), 500-531.

Lee, H. Y. and H. Hwang. 2016. Gradient of learnability in teaching English pronunciation to Korean learners. *The Journal of the Acoustical Society of America* 139(4), 1859-1872.

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

Levelt, W. J., A. Roelofs and A. S. Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22(1), 1-38.

Levis, J. M. 2005. Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly* 39(3), 369-377.

Levis, J. M. 2016. Research into practice: How research appears in pronunciation teaching materials. *Language Teaching* 49(3), 423-437.

Levis, J. M. 2020. Revisiting the intelligibility and nativeness principles. *Journal of Second Language Pronunciation* 6(3), 310-328.

Logan, J. S., S. E. Lively and D. B. Pisoni. 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.

Mahdi, H. S. and M. A. Mohsen. 2024. Enhancing Pronunciation Learning through High Variability Phonetic Training: A Meta-Analysis. *Language Teaching Research Quarterly* 40, 29-45.

Mora, J. C. 2005. Lexical knowledge effects on the discrimination of non-native phonemic contrasts in words and nonwords by Catalan/Spanish bilingual learners of English. In *Proceedings of the ISCA Workshop on Plasticity in Speech Perception,* 43-46.

Munro, M. J. and T. M. Derwing. 1995. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning* 45(1), 73-97.

Munro, M. J. and T. M. Derwing. 2015. A prospectus for pronunciation research in the 21st century: A point of view. *Journal of Second Language Pronunciation* 1(1), 11-42.

Munro, M. J. and T. M. Derwing. 2020. Foreign accent, comprehensibility and intelligibility, redux. *Journal of Second Language Pronunciation* 6(3), 283-309.

Nayar, P. B. 1997. ESL/EFL dichotomy today: Language politics or pragmatics? *TESOL Quarterly* 31(1), 9-37.

Ortega, M., J. C. Mora and I. Mora-Plaza. 2021. Differential effects of lexical and non-lexical high-variability phonetic training on the production of L2 vowels. In A. Kirkova-Naskova, A. Henderson, and J. Fouz-González, eds., *English Pronunciation Instruction: Research-based Insights,* 327-356. John Benjamins.

Page, M. J., J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, ... and D. Moher. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372(71), 1-9.

Pennington, M. C. 2021. Teaching pronunciation: The state of the art 2021. *RELC Journal* 52(1), 3-21.

Plonsky, L. and F. L. Oswald. 2014. How big is “big”? Interpreting effect sizes in L2 research. *Language Learning* 64(4), 878-912.

R Core Team 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. [Computer software]. Available online at https://www.R-project.org/

Rato, A. 2013. *Cross-Language Perception and Production of English Vowels by Portuguese Learners: The Effects of Perceptual Training.* Doctoral dissertation, Universidade do Minho.

Rato, A. and A. Carlet. 2020. Second language perception of English vowels by Portuguese learners: The effect of stimulus type. *Ilha do Desterro*, 73, 205-226.

Saito, H. and M. E. Ebsworth. 2004. Seeing English language teaching and learning through the eyes of Japanese EFL and ESL students. *Foreign Language Annals* 37(1), 111-124.

Sánchez-Meca, J. and F. Marín-Martínez. 2008. Confidence intervals for the overall effect size in random-

effects meta-analysis. *Psychological Methods* 13(1), 31-48.

Schwarzer, G., J. R. Carpenter, G. Rücker and M. Schumacher. 2024. *meta: General package for meta-analysis* (Version 6.5-0). [R package]. Available online at https://cran.r-project.org/web/packages/meta/meta.pdf

Sebastián-Gallés, N. and C. Baus. 2005. On the relationship between perception and production in L2 categories. In A. Cutler, ed., *Twenty-First Century Psycholinguistics: Four Cornerstones,* 266-277. Taylor and Francis Group. *ProQuest Ebook Central.*

Shinohara, Y. and P. Iverson. 2018. High variability identification and discrimination training for Japanese speakers learning English/r/-/l/. *Journal of Phonetics* 66, 242-251.

Shinohara, Y. and P. Iverson. 2021. The effect of age on English /r/-/l/ perceptual training outcomes for Japanese speakers. J*ournal of Phonetics* 89, 1-24.

Shum, K. K. M., T. K. F. Au, L. F. Romo and S.-A. Jun. 2021. Learning challenging L2 sounds via computer training: High-variability perceptual training for children and adults. *Language Learning and Development* 17(3), 327-342.

Tanner, M. and L. Henrichsen. 2022. Pronunciation in varied teaching and learning contexts. In J. M. Levis, T. M. Derwing, and S. Sonsaat-Hegelheimer, eds., *Second Language Pronunciation: Bridging the Gap between Research and Teaching,* 215-234. Wiley-Blackwell.

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.

Thomson, R. I. 2022a. Perception in pronunciation teaching. In J. M. Levis, T. M. Derwing, and S. Sonsaat-Hegelheimer, eds., *Second Language Pronunciation: Bridging the Gap between Research and Teaching,* 42-60. Wiley-Blackwell.

Thomson, R. I. 2022b. The relationship between L2 speech perception and production. In T. M. Derwing, M. J. Munro, and R. I. Thomson, eds., *The Routledge Handbook of Second Language Acquisition and Speaking*, 554-573. Routledge. ProQuest Ebook Central.

Thomson, R. I. and T. M. Derwing. 2015. The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics* 36(3), 326-344.

Thomson, R. I. and T. M. Derwing. 2016. Is phonemic training using nonsense or real words more effective? In *Proceedings of the 7th Pronunciation in Second Language Learning and Teaching Conference*, 88-97.

Tsukada, K., D. Birdsong, E. Bialystok, M. Mack, H. Sung and J. E. Flege. 2005. A developmental study of English vowel production and perception by native Korean adults and children. *Journal of Phonetics* 33(3), 263-290.

Uchihara, T., M. Karas and R. I. Thomson. 2024. Does perceptual high variability phonetic training improve L2 speech production? A meta-analysis of perception-production connection. *Applied Psycholinguistics* 45(4), 591-623.

Uchihara, T., M. Karas and R. I. Thomson. 2025. High variability phonetic training (HVPT): A meta-analysis of L2 perceptual training studies. *Studies in Second Language Acquisition*, 1-34.

Viechtbauer, W. 2005. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics* 30(3), 261-293.

Wayland, R. P. 2007. The relationship between identification and discrimination in cross-language perception: The case of Korean and Thai. In O.-S. Bohn and M. J. Munro, eds., *Language Experience in Second Language Speech Learning: In Honor of James E. Flege*, 201-218. John Benjamins.

Wong, 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*, Interspeech*, 524-528.
World Health Organization. (n.d.). *Adolescent health*. WHO. Retrieved June 21, 2025, Available online at https://www.who.int/health-topics/adolescent-health
Zhang, X., B. Cheng and Y. Zhang. 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.

Examples in: English
Applicable Languages: English
Applicable Level: ALL

Appendix A. Keywords Used for Database Searches (*Publish or Perish*)

hvpt, high variability phonetic training, training, l2, second language, pronunciation, perception, production, phonetic training, auditory training, stimulus variability, perceptual learning, talker variability

Appendix B. Included Studies in the Meta-Analysis

Aliaga-García, 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.
Aliaga-García, C. and J. C. Mora. 2009. Assessing the effects of phonetic training on L2 sound perception and production. In M. A. Watkins, A. S. Rauber, and B. O. Baptista, eds., *Recent Research in Second Language Phonetics/Phonology: Perception and Production*, 2-31. Cambridge Scholars Publishing.
*Alves, U. K. and P. L. Luchini. 2017. Effects of perceptual training on the identification and production of word-initial voiceless stops by Argentinean learners of English. *Ilha Do Desterro* 70(3), 15-32.
Carlet, A. and J. Cebrian. 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. B. Jespersen, and M. H. Sørensen, eds., *A Sound Approach to Language Matters – In Honor of Ocke-Schwen Bohn,* 91-119. Dept. of English, School of Communication and Culture, Aarhus University.
Carlet, A. and J. Cebrian. 2022. The roles of task, segment type, and attention in L2 perceptual training. *Applied Psycholinguistics* 43(2), 271-299.
Carlet, A., J. Cebrian, N. Gavaldà and C. Gorba. 2022. Does metalinguistic knowledge about the L2 enhance the effectiveness of L2 perceptual training? In B. Blecua, J. Cicres, M. Espejel, and M. J. Machuca, eds., *Propuestas en Fonética Experimental: Enfoques Metodológicos y Nuevas Tecnologías*, 31-35. Universitat de Girona-Servei de Publicacions.
Fouz-González, J. and J. A. Mompean. 2021. Exploring the potential of phonetic symbols and keywords as labels for perceptual training. *Studies in Second Language Acquisition* 43(2), 297-328.
Georgiou, G. P. 2022. The impact of auditory perceptual training on the perception and production of English vowels by Cypriot Greek children and adults. *Language Learning and Development* 18(4), 379-392.
*Huang, Y. F. 2013. *The Effects of Two Methods on Training EFL University Students in Taiwan to Identify Three Non-Native Phonetic Contrasts,* Doctoral dissertation, The Ohio State University.
Lacabex, E. G. and F. Gallardo del Puerto. 2014. Two phonetic-training procedures for young learners: Investigating instructional effects on perceptual awareness. *The Canadian Modern Language Review* 70(4), 500-531.

Lambacher, S., W. Martens, K. Kakehi, C. A. Marasinghe and G. Molholt. 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.

Li, Y. 2015. *Audio-Visual Training Effect on L2 Perception and Production of English /θ/-/s/ and /ð/-/z/ by Mandarin Speakers,* Doctoral dissertation, Newcastle University.

Rato, A. A. dos S. 2013. *Cross-Language Perception and Production of English Vowels by Portuguese Learners: The Effects of Perceptual Training,* Doctoral dissertation, Universidade do Minho.

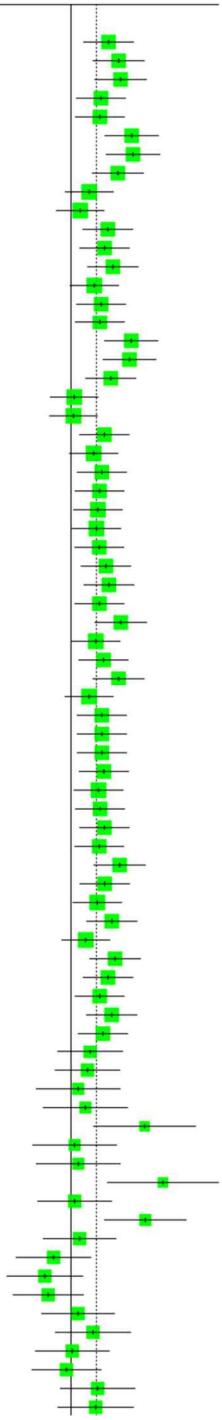
Shum, K. K. M., T. K. F. Au, L. F. Romo and S.-A. Jun. 2021. Learning challenging L2 sounds via computer training: High-variability perceptual training for children and adults. *Language Learning and Development* 17(3), 327-342.

Ueda, R. and K. Hashimoto. 2018. Perceptual training in a classroom setting: Phonemic category formation by Japanese EFL learners. In *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference*, 237-249.

Note. Studies marked with * were identified through manual searches of the reference lists of four previous meta-analyses.

Appendix C. Forest Plots of HVPT Effects by Study Outcomes

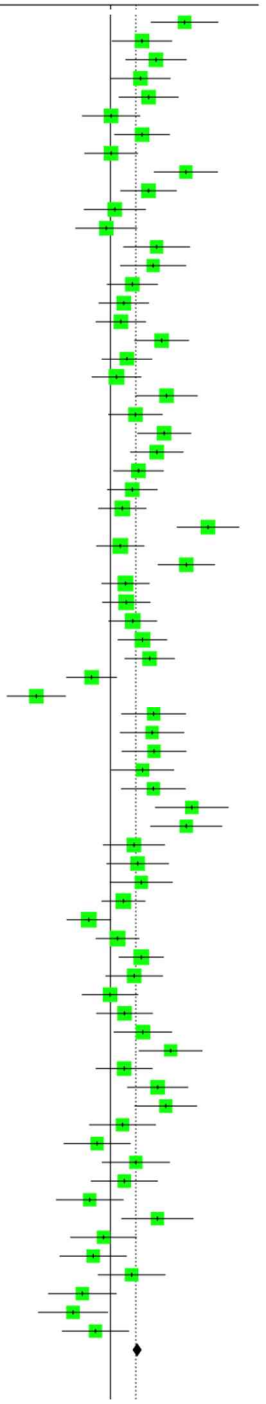
Study or Subgroup	Experimental Mean	SD	Total	Control Mean	SD	Total	Weight	Std. Mean Difference IV, Random, 95% CI
subgroup = posttest								
Aliaga-Garcia (a) 2017	76.17	10.5900	32	61.88	22.2800	20	0.6%	0.88 [0.29; 1.46]
Aliaga-Garcia (b) 2017	80.08	11.8200	32	62.50	20.2800	20	0.6%	1.11 [0.51; 1.71]
Aliaga-Garcia (c) 2017	97.79	4.3600	32	90.00	9.2100	20	0.6%	1.16 [0.55; 1.76]
Aliaga-Garcia (d) 2017	93.88	8.7900	32	86.04	13.9400	20	0.6%	0.70 [0.12; 1.28]
Aliaga-Garcia (e) 2017	76.17	16.7700	32	64.58	17.2000	20	0.6%	0.67 [0.10; 1.25]
Aliaga-Garcia (f) 2017	78.64	13.7500	32	57.29	16.4900	20	0.6%	1.42 [0.79; 2.04]
Aliaga-Garcia (g) 2017	83.72	14.2900	32	62.92	13.9500	20	0.6%	1.45 [0.82; 2.08]
Aliaga-Garcia (h) 2017	76.04	11.9300	32	60.21	17.3900	20	0.6%	1.09 [0.49; 1.69]
Aliaga-Garcia (i) 2017	71.48	11.0500	32	67.08	8.5400	20	0.7%	0.43 [-0.14; 0.99]
Aliaga-Garcia (j) 2017	57.16	10.2700	32	54.58	14.0400	20	0.7%	0.21 [-0.35; 0.77]
Aliaga-Garcia (k) 2017	67.06	12.3100	32	56.46	11.8800	20	0.6%	0.86 [0.27; 1.44]
Aliaga-Garcia (l) 2017	75.00	11.6400	32	61.88	22.2800	20	0.6%	0.78 [0.20; 1.36]
Aliaga-Garcia (m) 2017	80.08	15.9400	32	62.50	20.2800	20	0.6%	0.98 [0.39; 1.57]
Aliaga-Garcia (n) 2017	94.66	7.9400	32	90.00	9.2100	20	0.7%	0.54 [-0.03; 1.11]
Aliaga-Garcia (o) 2017	94.14	9.3900	32	86.04	13.9400	20	0.6%	0.70 [0.13; 1.28]
Aliaga-Garcia (p) 2017	78.52	22.1300	32	64.58	17.2000	20	0.6%	0.67 [0.10; 1.25]
Aliaga-Garcia (q) 2017	80.60	16.2400	32	57.29	16.4900	20	0.6%	1.41 [0.78; 2.03]
Aliaga-Garcia (r) 2017	83.07	14.8900	32	62.92	13.9500	20	0.6%	1.36 [0.74; 1.99]
Aliaga-Garcia (s) 2017	72.53	9.5200	32	60.21	17.3900	20	0.6%	0.93 [0.34; 1.52]
Aliaga-Garcia (t) 2017	67.84	10.7500	32	67.08	8.5400	20	0.7%	0.08 [-0.48; 0.63]
Aliaga-Garcia (u) 2017	55.34	12.2600	32	54.58	14.0400	20	0.7%	0.06 [-0.50; 0.62]
Aliaga-Garcia (v) 2017	66.15	12.4700	32	56.46	11.8800	20	0.6%	0.78 [0.20; 1.36]
Aliaga-Garcia (w) 2017	73.96	12.6500	32	63.09	28.6300	20	0.7%	0.53 [-0.04; 1.10]
Aliaga-Garcia (x) 2017	92.71	8.4000	32	76.19	35.0300	20	0.6%	0.72 [0.14; 1.30]
Aliaga-Garcia (y) 2017	92.71	10.3100	32	77.38	34.3500	20	0.6%	0.67 [0.09; 1.24]
Aliaga-Garcia (z) 2017	97.92	5.6000	32	83.33	36.4000	20	0.7%	0.63 [0.06; 1.20]
Aliaga-Garcia (za) 2017	89.58	13.8800	32	75.00	35.0100	20	0.7%	0.59 [0.02; 1.16]
Aliaga-Garcia (zb) 2017	99.48	2.9500	32	84.52	36.0800	20	0.6%	0.66 [0.08; 1.23]
Aliaga-Garcia (zc) 2017	90.62	12.6500	32	71.43	33.9300	20	0.6%	0.82 [0.23; 1.40]
Aliaga-Garcia (zd) 2017	96.87	11.4800	32	75.71	35.2400	20	0.6%	0.89 [0.30; 1.47]
Aliaga-Garcia (ze) 2017	97.39	7.4600	32	82.14	35.4800	20	0.6%	0.66 [0.09; 1.24]
Aliaga-Garcia (zf) 2017	98.44	4.9300	32	72.62	34.9600	20	0.6%	1.16 [0.56; 1.77]
Aliaga-Garcia (zg) 2017	67.41	18.9400	32	53.06	31.4100	20	0.7%	0.58 [0.01; 1.15]
Aliaga-Garcia (zh) 2017	67.50	21.4000	32	48.57	29.0500	20	0.6%	0.76 [0.18; 1.34]
Aliaga-Garcia (zi) 2017	86.46	16.6300	32	60.71	30.3900	20	0.6%	1.11 [0.51; 1.71]
Aliaga-Garcia (zj) 2017	84.89	13.6200	32	75.00	33.1700	20	0.7%	0.42 [-0.14; 0.99]
Aliaga-Garcia (zk) 2017	93.30	12.0300	32	76.53	33.9000	20	0.6%	0.72 [0.14; 1.30]
Aliaga-Garcia (zl) 2017	96.35	8.1800	32	79.76	35.3100	20	0.6%	0.72 [0.14; 1.30]
Aliaga-Garcia (zm) 2017	77.08	9.2300	32	63.09	28.6300	20	0.6%	0.72 [0.14; 1.30]
Aliaga-Garcia (zn) 2017	93.75	8.2000	32	76.19	35.0300	20	0.6%	0.77 [0.19; 1.35]
Aliaga-Garcia (zo) 2017	92.19	10.3600	32	77.38	34.3500	20	0.7%	0.64 [0.07; 1.22]
Aliaga-Garcia (zp) 2017	98.93	4.1000	32	83.33	36.4000	20	0.6%	0.68 [0.10; 1.25]
Aliaga-Garcia (zq) 2017	93.23	10.2500	32	75.00	35.0100	20	0.6%	0.78 [0.20; 1.36]
Aliaga-Garcia (zr) 2017	99.48	2.9500	32	84.52	36.0800	20	0.6%	0.66 [0.08; 1.23]
Aliaga-Garcia (zs) 2017	96.35	7.0000	32	71.43	33.9300	20	0.6%	1.13 [0.53; 1.74]
Aliaga-Garcia (zt) 2017	95.00	13.4400	32	75.71	35.2400	20	0.6%	0.79 [0.21; 1.37]
Aliaga-Garcia (zu) 2017	96.35	8.1800	32	82.14	35.4800	20	0.7%	0.61 [0.04; 1.19]
Aliaga-Garcia (zv) 2017	94.79	9.8700	32	72.62	34.9600	20	0.6%	0.95 [0.36; 1.54]
Aliaga-Garcia (zw) 2017	62.50	24.0000	32	53.06	31.4100	20	0.7%	0.34 [-0.22; 0.91]
Aliaga-Garcia (zx) 2017	72.50	18.1400	32	48.57	29.0500	20	0.6%	1.03 [0.43; 1.62]
Aliaga-Garcia (zy) 2017	81.25	17.8300	32	60.71	30.3900	20	0.6%	0.86 [0.28; 1.45]
Aliaga-Garcia (zz) 2017	90.62	13.3400	32	75.00	33.1700	20	0.6%	0.67 [0.09; 1.24]
Aliaga-Garcia (zza) 2017	97.32	6.7300	32	76.53	33.9000	20	0.6%	0.95 [0.36; 1.54]
Aliaga-Garcia (zzb) 2017	96.87	7.8500	32	79.76	35.3100	20	0.6%	0.74 [0.17; 1.32]
Aliaga-Garcia & Mora (a) 2009	73.57	9.6300	18	69.15	9.5600	11	0.5%	0.45 [-0.31; 1.21]
Aliaga-Garcia & Mora (b) 2009	84.90	11.1900	18	80.89	8.0600	11	0.5%	0.38 [-0.37; 1.14]
Alves & Luchini (a) 2017	98.96	11.8800	8	96.88	11.6200	8	0.4%	0.17 [-0.81; 1.15]
Alves & Luchini (b) 2017	100.00	12.0000	8	95.83	11.5000	8	0.4%	0.34 [-0.65; 1.32]
Alves & Luchini (c) 2017	89.58	10.7500	8	71.88	8.6200	8	0.3%	1.72 [0.52; 2.91]
Alves & Luchini (g) 2017	97.92	11.7500	8	96.88	11.6200	8	0.4%	0.08 [-0.90; 1.06]
Alves & Luchini (h) 2017	97.92	11.7500	8	95.83	11.5000	8	0.4%	0.17 [-0.81; 1.15]
Alves & Luchini (i) 2017	94.79	11.3800	8	71.88	8.6200	8	0.3%	2.15 [0.85; 3.44]
Carlet & Cebrian (a) 2022	62.80	9.4000	12	61.90	11.1000	9	0.5%	0.09 [-0.78; 0.95]
Carlet & Cebrian (b) 2022	79.70	9.3000	17	61.90	11.1000	9	0.4%	1.73 [0.78; 2.69]
Carlet & Cebrian (c) 2022	64.00	9.3000	13	61.90	11.1000	9	0.5%	0.20 [-0.65; 1.05]
Carlet & Cebrian (d) 2022	57.30	10.6000	12	61.90	11.1000	9	0.5%	-0.41 [-1.28; 0.47]
Carlet & Cebrian (e) 2022	74.30	11.7000	12	82.60	14.7000	9	0.5%	-0.61 [-1.50; 0.28]
Carlet & Cebrian (f) 2022	75.40	12.2000	17	82.60	14.7000	9	0.5%	-0.53 [-1.36; 0.29]
Carlet & Cebrian (g) 2022	84.60	9.3000	13	82.60	14.7000	9	0.5%	0.16 [-0.69; 1.02]
Carlet & Cebrian (h) 2022	88.40	6.8000	12	82.60	14.7000	9	0.5%	0.51 [-0.37; 1.39]
Carlet & Cebrian (i) 2022	71.70	6.6000	12	71.50	6.4000	9	0.5%	0.03 [-0.83; 0.89]
Carlet & Cebrian (j) 2022	70.60	9.0000	17	71.50	6.4000	9	0.5%	-0.11 [-0.91; 0.70]
Carlet & Cebrian (k) 2022	75.60	6.3000	13	71.50	6.4000	9	0.5%	0.62 [-0.25; 1.50]
Carlet & Cebrian (l) 2022	75.20	6.0000	12	71.50	6.4000	9	0.5%	0.58 [-0.31; 1.46]



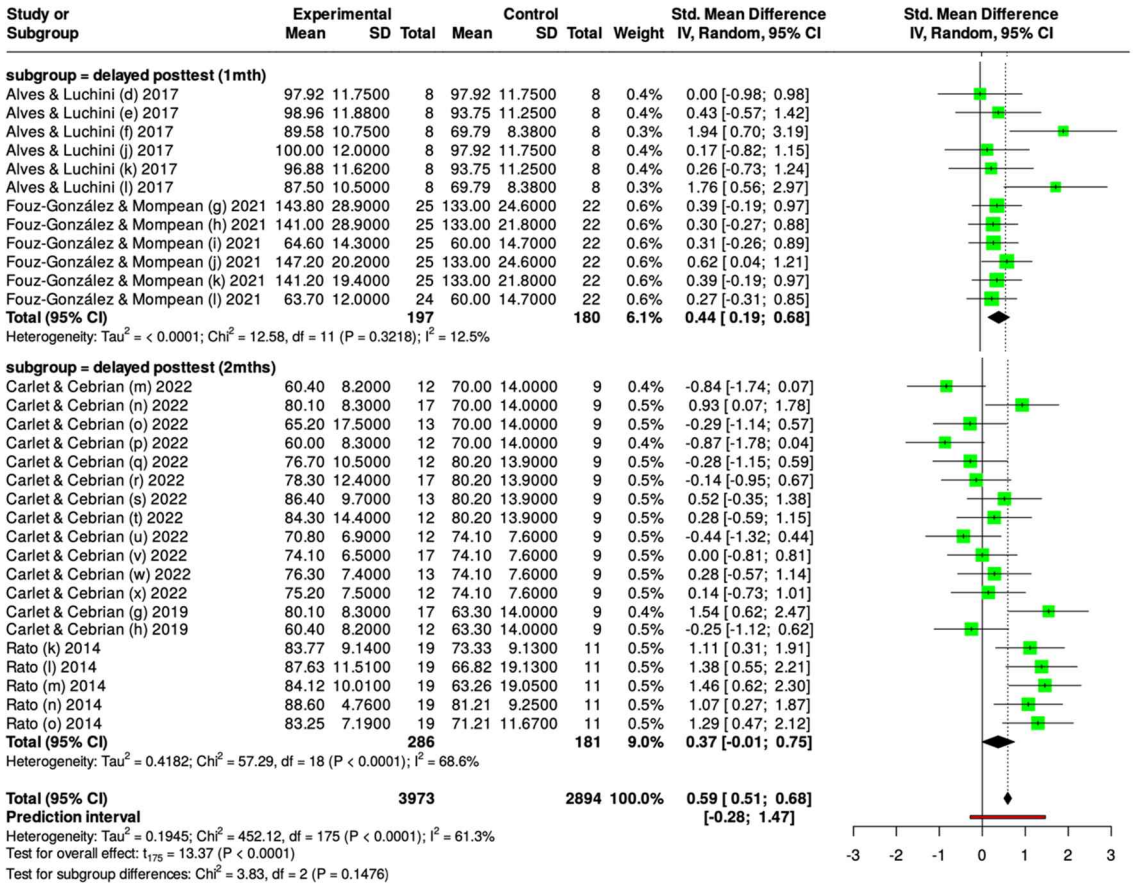
(Continued)

Study or Subgroup	Experimental Mean	SD	Total	Control Mean	SD	Total	Weight	Std. Mean Difference IV, Random, 95% CI
Carlet & Cebrian (a) 2019	79.10	13.3000	20	57.80	10.2000	16	0.5%	1.73 [0.95; 2.51]
Carlet & Cebrian (b) 2019	65.30	9.7000	18	57.80	10.2000	16	0.6%	0.74 [0.04; 1.44]
Carlet & Cebrian (c) 2019	80.40	9.8000	20	68.40	12.4000	16	0.6%	1.06 [0.36; 1.77]
Carlet & Cebrian (d) 2019	75.90	8.3000	18	68.40	12.4000	16	0.6%	0.70 [0.01; 1.40]
Carlet & Cebrian (e) 2019	88.50	9.5000	20	79.50	10.3000	16	0.6%	0.89 [0.20; 1.58]
Carlet & Cebrian (f) 2019	79.70	11.1000	18	79.50	10.3000	16	0.6%	0.02 [-0.66; 0.69]
Carlet et al. (a) 2022	65.30	9.7000	20	57.80	10.2000	20	0.6%	0.74 [0.10; 1.38]
Carlet et al. (b) 2022	79.70	11.1000	20	79.50	10.3000	20	0.6%	0.02 [-0.60; 0.64]
Carlet et al. (c) 2022	79.10	13.3000	20	57.80	10.2000	20	0.5%	1.76 [1.02; 2.50]
Carlet et al. (d) 2022	88.50	9.5000	20	79.50	10.3000	20	0.6%	0.89 [0.24; 1.54]
Carlet et al. (e) 2022	59.00	15.0000	15	57.50	13.0000	15	0.6%	0.10 [-0.61; 0.82]
Carlet et al. (f) 2022	65.10	16.0000	15	66.50	12.0000	15	0.6%	-0.10 [-0.81; 0.62]
Carlet et al. (g) 2022	69.90	9.0000	15	57.50	13.0000	15	0.5%	1.08 [0.31; 1.85]
Carlet et al. (h) 2022	78.30	11.0000	15	66.50	12.0000	15	0.5%	1.00 [0.23; 1.76]
Fouz-González & Mompean (a) 2021	140.80	28.2000	24	130.10	3.2100	22	0.6%	0.51 [-0.08; 1.10]
Fouz-González & Mompean (b) 2021	140.30	27.6000	24	132.60	19.8000	22	0.6%	0.31 [-0.27; 0.90]
Fouz-González & Mompean (c) 2021	63.80	15.3000	24	60.40	11.7000	22	0.6%	0.24 [-0.34; 0.82]
Fouz-González & Mompean (d) 2021	146.60	18.5000	24	130.10	3.2100	22	0.6%	1.20 [0.56; 1.83]
Fouz-González & Mompean (e) 2021	140.50	20.4000	24	132.60	19.8000	22	0.6%	0.39 [-0.20; 0.97]
Fouz-González & Mompean (f) 2021	62.10	11.7000	25	60.40	11.7000	22	0.7%	0.14 [-0.43; 0.72]
Georgiou (a) 2022	71.00	11.3000	21	56.00	11.1000	16	0.6%	1.31 [0.59; 2.03]
Georgiou (b) 2022	64.00	14.3000	22	55.00	16.0000	19	0.6%	0.58 [-0.04; 1.21]
Huang (a) 2013	67.93	15.5800	24	50.60	11.2400	24	0.6%	1.25 [0.63; 1.88]
Huang (b) 2013	64.11	13.2400	23	50.60	11.2400	24	0.6%	1.08 [0.47; 1.70]
Huang (c) 2013	71.11	12.9700	24	61.62	15.3900	24	0.6%	0.66 [0.07; 1.24]
Huang (d) 2013	68.72	11.5300	23	61.62	15.3900	24	0.6%	0.51 [-0.07; 1.09]
Lacabex & Gallardo (a) 2014	80.79	12.6000	25	77.07	13.8000	25	0.7%	0.28 [-0.28; 0.83]
Lacabex & Gallardo (b) 2014	66.39	20.9000	25	22.62	16.7000	25	0.6%	2.28 [1.55; 3.00]
Lacabex & Gallardo (c) 2014	79.82	9.2000	25	77.07	13.8000	25	0.7%	0.23 [-0.33; 0.79]
Lacabex & Gallardo (d) 2014	59.30	23.5000	25	22.62	16.7000	25	0.6%	1.77 [1.11; 2.43]
Lambacher et al. (a) 2005	68.40	31.9000	34	56.50	35.4000	20	0.7%	0.35 [-0.20; 0.91]
Lambacher et al. (b) 2005	59.20	32.6000	34	46.80	33.2000	20	0.7%	0.37 [-0.18; 0.93]
Lambacher et al. (c) 2005	66.50	35.1000	34	47.10	39.1000	20	0.7%	0.52 [-0.04; 1.08]
Lambacher et al. (d) 2005	70.80	29.9000	34	46.70	34.7000	20	0.7%	0.75 [0.18; 1.32]
Lambacher et al. (e) 2005	85.60	25.8000	34	57.90	35.7000	20	0.6%	0.92 [0.34; 1.50]
Li (a) 2015	3.09	0.9300	27	3.43	0.4400	20	0.6%	-0.44 [-1.02; 0.15]
Li (b) 2015	2.63	0.5900	27	3.46	0.2300	20	0.6%	-1.73 [-2.41; -1.04]
Rato (a) 2014	81.44	8.6700	22	71.81	10.4500	12	0.5%	1.01 [0.26; 1.76]
Rato (b) 2014	83.71	13.6800	22	69.17	16.1500	12	0.5%	0.97 [0.23; 1.72]
Rato (c) 2014	77.42	12.4800	22	59.58	23.5600	12	0.5%	1.02 [0.27; 1.77]
Rato (d) 2014	85.76	5.0100	22	80.56	9.2200	12	0.5%	0.75 [0.02; 1.48]
Rato (e) 2014	83.12	7.0000	22	73.47	12.7800	12	0.5%	1.00 [0.26; 1.75]
Rato (f) 2014	90.28	6.7900	22	72.45	12.5000	12	0.5%	1.90 [1.05; 2.75]
Rato (g) 2014	89.27	8.2800	22	68.06	16.3500	12	0.5%	1.77 [0.94; 2.60]
Rato (h) 2014	67.42	16.8900	22	57.17	20.3600	12	0.6%	0.55 [-0.17; 1.27]
Rato (i) 2014	75.63	11.8100	22	66.90	15.9600	12	0.6%	0.64 [-0.08; 1.36]
Rato (j) 2014	70.08	15.1200	22	58.56	16.4300	12	0.5%	0.72 [-0.00; 1.45]
Shum et al. (a) 2021	65.60	7.6000	29	63.80	3.5000	32	0.7%	0.31 [-0.20; 0.81]
Shum et al. (b) 2021	86.30	12.6000	29	91.80	8.9000	32	0.7%	-0.50 [-1.01; 0.01]
Shum et al. (c) 2021	54.70	7.1000	29	53.70	4.6000	32	0.7%	0.17 [-0.34; 0.67]
Shum et al. (d) 2021	66.10	10.6000	29	59.90	6.0000	32	0.7%	0.72 [0.20; 1.24]
Shum et al. (e) 2021	66.30	6.9000	23	62.80	4.8000	15	0.6%	0.56 [-0.11; 1.22]
Shum et al. (f) 2021	90.30	12.9000	23	90.40	12.7000	15	0.6%	-0.01 [-0.66; 0.64]
Shum et al. (g) 2021	57.30	7.3000	23	55.20	4.1000	15	0.6%	0.33 [-0.33; 0.98]
Shum et al. (h) 2021	63.30	9.0000	23	56.70	7.7000	15	0.6%	0.76 [0.08; 1.43]
Shum et al. (i) 2021	86.10	7.3000	18	76.30	6.3000	18	0.5%	1.41 [0.67; 2.14]
Shum et al. (j) 2021	99.30	2.4000	18	97.20	8.7000	18	0.6%	0.32 [-0.34; 0.98]
Shum et al. (k) 2021	75.80	10.9000	18	65.20	7.6000	18	0.6%	1.10 [0.40; 1.81]
Shum et al. (l) 2021	89.70	9.2000	18	77.00	10.0000	18	0.5%	1.29 [0.57; 2.02]
Ueda & Hashimoto (a) 2019	13.38	1.6100	13	12.85	1.9900	13	0.5%	0.28 [-0.49; 1.06]
Ueda & Hashimoto (b) 2019	8.85	2.4800	13	9.69	2.7800	13	0.5%	-0.31 [-1.08; 0.47]
Ueda & Hashimoto (c) 2019	13.92	1.3200	13	12.92	1.8900	13	0.5%	0.59 [-0.19; 1.38]
Ueda & Hashimoto (d) 2019	10.00	2.8800	13	9.23	1.4800	13	0.5%	0.33 [-0.45; 1.10]
Ueda & Hashimoto (e) 2019	8.69	2.9700	13	10.15	2.9000	13	0.5%	-0.48 [-1.26; 0.30]
Ueda & Hashimoto (f) 2019	14.38	1.0800	13	12.69	1.8100	13	0.5%	1.10 [0.26; 1.93]
Ueda & Hashimoto (g) 2019	2.62	2.4500	13	2.92	0.7300	13	0.5%	-0.16 [-0.93; 0.61]
Ueda & Hashimoto (h) 2019	1.92	1.1400	13	2.38	1.0800	13	0.5%	-0.40 [-1.18; 0.38]
Ueda & Hashimoto (i) 2019	3.54	0.6300	13	3.23	0.5800	13	0.5%	0.50 [-0.29; 1.28]
Ueda & Hashimoto (j) 2019	2.46	0.6300	13	2.92	0.7300	13	0.5%	-0.65 [-1.45; 0.14]
Ueda & Hashimoto (k) 2019	2.00	1.1800	13	3.00	1.0400	13	0.5%	-0.87 [-1.68; -0.06]
Ueda & Hashimoto (l) 2019	3.15	0.6600	13	3.38	0.6200	13	0.5%	-0.35 [-1.12; 0.43]
Total (95% CI)			**3490**			**2533**	**84.9%**	**0.62 [0.53; 0.72]**

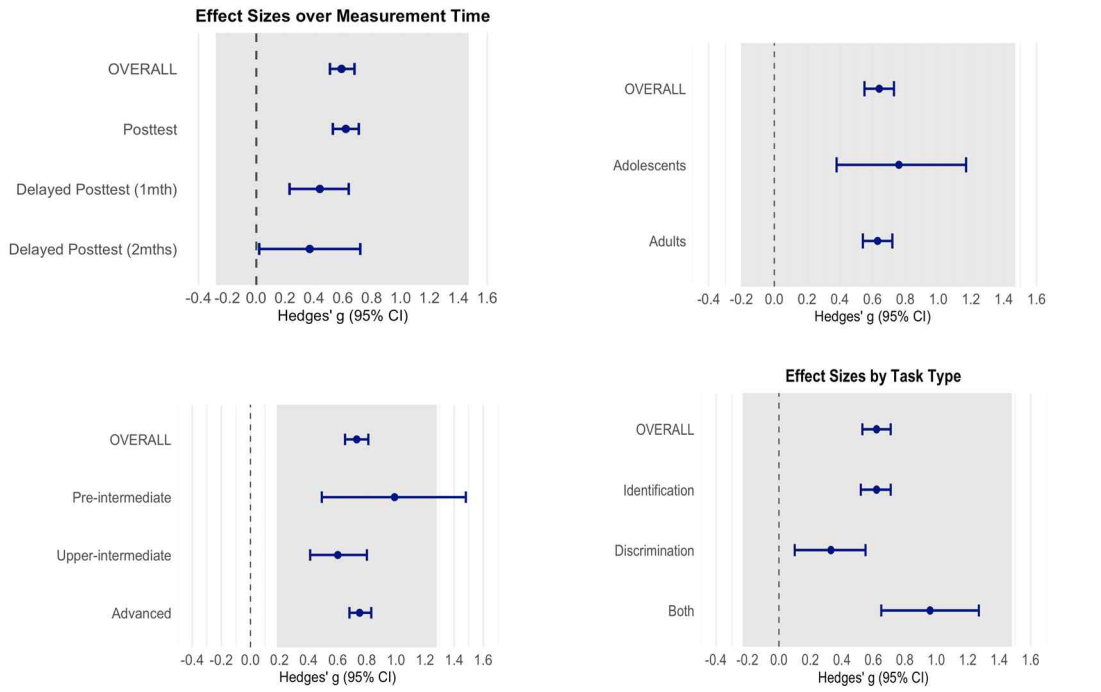
Heterogeneity: $Tau^2 = 0.1837$; $Chi^2 = 374.63$, df = 144 ($P < 0.0001$); $I^2 = 61.6\%$



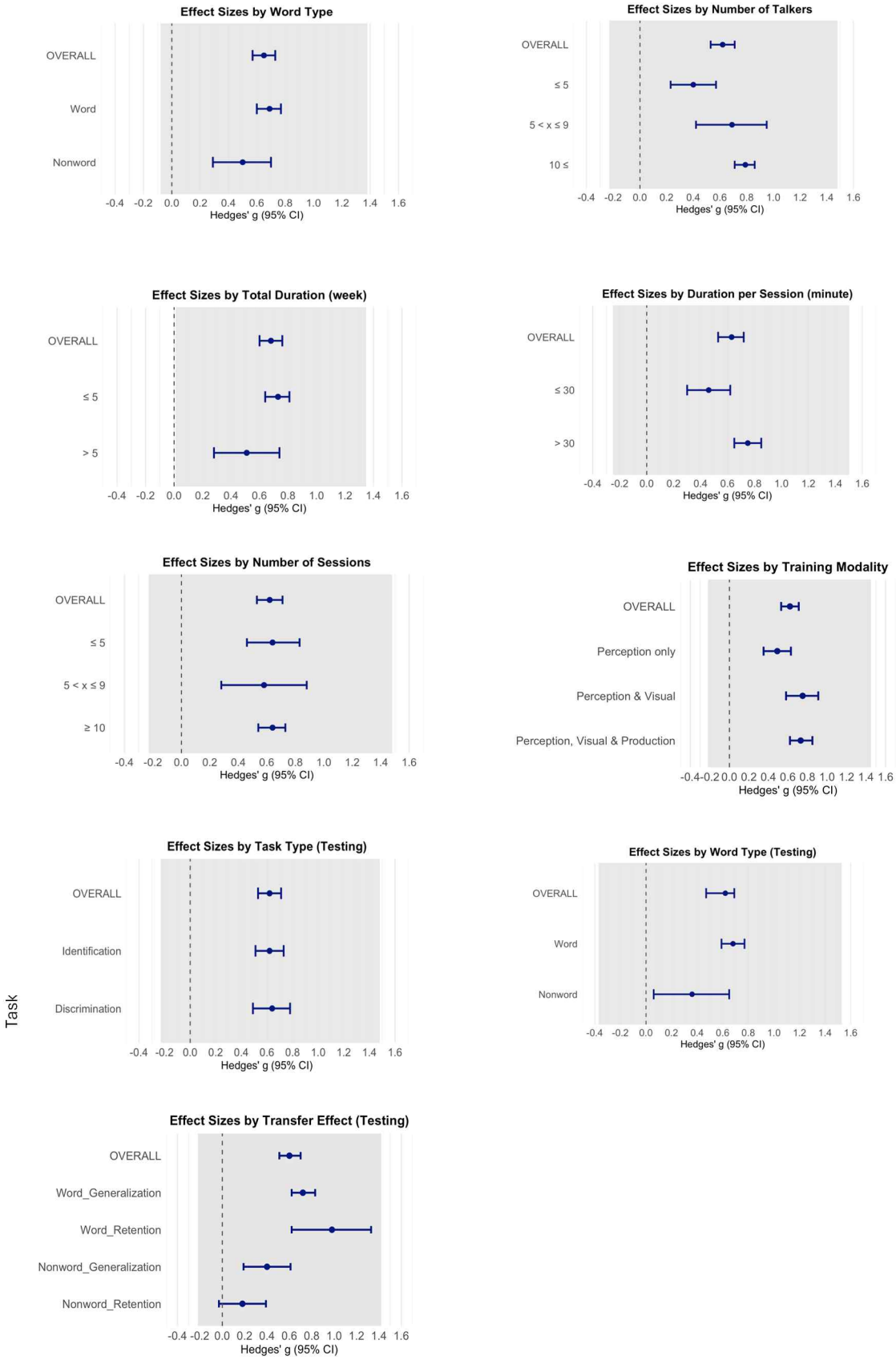
(Continued)



Appendix D. Forest Plots of HVPT Effects by Moderating Variables



(Continued)



Note. The gray-shaded area represents the overall prediction interval (PI).