CAMBRIDGE
UNIVERSITY PRESS

ORIGINAL ARTICLE

Does perceptual high variability phonetic training improve L2 speech production? A meta-analysis of perception-production connection

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

[1]Graduate School of International Cultural Studies, Tohoku University, Sendai 980-8576, Japan and [2]Department of Applied Linguistics, Brock University, St Catharines, Canada
Corresponding author: Takumi Uchihara; Email: takumi@tohoku.ac.jp

Abstract

This meta-analysis of 31 studies aimed to determine the effectiveness of perception-based high variability phonetic training (HVPT) for second language (L2) production learning and to identify learner-related and methodological variables that influence production gains. Based on independent effect sizes for 43 within-participant and 17 between-participant designs, small-to-medium effects of post-training improvement were found. The average production gains for trained items and untrained items were 10.50% and 4.50%, respectively. Neither strong support for long-term retention of production learning nor generalization to untrained stimuli was observed, however. Moderator analyses showed that post-training production gains were influenced by a number of factors related to learner profiles (age and learning context), training features (provision of phonetic information, training duration, and training time per session), and features of production tests (elicitation tasks, prompt modality, and outcome measures). The relationship between perception and production gains was negligible at the participant level, but was significant and moderate at the level of individual studies for post-training and retention data. These findings provide partial support for a perception-production link. This study makes several recommendations for future studies investigating the effects of HVPT on L2 speech production learning.

Keywords: high variability phonetic training; L2 speech production; meta-analysis; perception-production link

Introduction

The efficacy of high variability phonetic training (HVPT), an increasingly popular perceptual training technique, has been widely documented in the literature on second language (L2) speech learning (Barriuso & Hayes-Harb, 2018; Thomson, 2018). In HVPT, L2 learners are trained to perceive target sounds, which can include speech segments (vowels and consonants) and/or suprasegmentals

(e.g., lexical tone), produced by multiple speakers, in varied phonetic contexts. In a seminal work by Logan et al. (1991), Japanese native speakers of L2 English heard minimal pair words contrasting /l/ and /r/ (e.g., lead vs. read), indicated which sound they heard using a two-alternative forced-choice identification task, and received trial-by-trial corrective feedback. As a result of training, their perception accuracy improved significantly from pretest (78.1%) to posttest (85.9%). Subsequent studies have since confirmed that perception accuracy of many other L2 sounds improves through similar training, including vowels (Lambacher et al., 2005; Thomson, 2012b), stops (Flege, 1995b), fricatives (Lengeris & Nicolaidis, 2014), tones (Wang et al., 1999), and syllable structures (Huensch & Tremblay, 2015). The benefit of HVPT has also been shown to generalize to improvement in the perception of trained target sounds in new words/phonetic contexts (Carlet & Cebrian, 2022) and those produced by unfamiliar talkers (Herd et al., 2013). Further, gains have been retained for 2 weeks (Lee & Lyster, 2016) to 6 months (Silpachai, 2020). A recent meta-analysis conducted by Uchihara et al. (2021) confirmed that HVPT led to large immediate gains in perception (g = 0.96 and 0.97), to their generalization to untrained stimuli, and to long-term retention of gains in perception.

This area of research is important not only because of its practical implications for learners but also because it has provided theoretical insights into the relationship between speech perception and speech production. HVPT research has often found that perceptual training can lead to improvements in the production of the trained sounds, even in the absence of an explicit focus on production during training (e.g., Bradlow et al., 1997, 1999). The positive effect of perception-only training on production aligns with Flege's (1995a) Speech Learning Model (SLM), which postulates that the accuracy of reliable L2 perceptual representations determines the degree to which L2 sounds are produced accurately (for a revised model endorsing the bidirectional perception-production link, see Flege & Bohn, 2021). Nevertheless, the perception-production link seems to vary considerably across studies, ranging from 0-1% (Hwang & Lee, 2015), 1-7% (Hazan et al., 2005), 4% (Iverson et al., 2012), 6% (Bradlow et al., 1999), 8% (Lambacher et al., 2005), 18% (Wang et al., 2003), to 21-25% (Okuno & Hardison, 2016). Similarly, correlations between perception and production gains also vary substantially across studies, ranging from .00 (Bradlow et al., 1999), .04 (Hwang & Lee, 2015), .19 (Ghaffarvand Mokari & Werner, 2018), .32 (Wong, 2013), .42 (Carlet, 2017) to .55 (Yang et al., 2021). These variations may stem from different characteristics of learners (e.g., age of learning, length of residence, learning contexts) and different methodological procedures (e.g., training duration, provision of phonetic information) adopted by each study (Melnik-Leroy et al., 2022; Nagle & Baese-Berk, 2022; Thomson, 2022a).

To date, two meta-analyses are available (Sakai & Moorman, 2018; Zhang et al., 2021). While both provide insights into the source of variation in learning outcomes as well as the overall effectiveness of HVPT in improving speech production, they are limited to the examination of a specific component of HVPT (i.e., focusing on a single component of HVPT or multi-talker variability in Zhang et al., 2021) or a broader conceptualization of perceptual training (i.e., focusing on various training procedures including but not limited to HVPT in Sakai & Moorman, 2018). In response to a recent explosion of HVPT-specific research, including many studies

since 2018, and the call for real-world application of this specific technique (Thomson, 2018), the time is ripe for a meta-analysis to examine the overall efficacy of HVPT and determine ways of maximizing its potential for promoting perception-to-production transfer. Thus, defining HVPT in terms of three key features (talker variability, phonetic context variability, and corrective feedback), the current meta-analysis aims to determine the extent to which perceptual HVPT impacts production learning and to identify the learner-related and methodological variables that contribute to between-study variation in the production transfer.

The relationship between L2 speech perception and production

Flege's (1995a) highly influential SLM makes several claims concerning the relationship between L2 perception and production. The SLM postulates that accurate perceptual representations of L2 sounds inform accurate L2 speech production. A key condition predicting successful L2 pronunciation is that learners can discriminate phonetic differences that exist between sounds found in their L1 phonological inventory and similar sounds in the L2. When learners can discern perceptual differences between crosslinguistically similar sounds, a new category for the L2 will be formed. After the establishment of a new category and with continued exposure to L2 input, the SLM predicts that learners will eventually be able to produce the target sound accurately. In contrast, in a recently revised model (SLM-r), Flege and Bohn (2021) propose a bidirectional, co-evolving perception-production relationship, adding greater nuance to the perception-first view of L2 speech development proposed by Flege's (1995a) SLM. This revised view is more consistent with Nearey's (1997) Double-Weak Theory of L1 speech perception and production, which argues for autonomous yet cooperating perception and production systems.

Existing research almost exclusively supports Flege's (1995a) original claim that improvements in production follow improvements in perception, often after a time lag between the two. Cross-sectional studies have suggested a positive correlation between perception and production accuracy (Flege et al., 1997, 1999; Baker & Trofimovich, 2006; Jia et al., 2006; Thomson, 2008; Hattori & Iverson, 2009; Saito & van Poeteren, 2018; Melnik-Leroy et al., 2022). However, several studies have also reported nonsignificant correlations (Peperkamp & Bouchon, 2011; Kartushina & Frauenfelder, 2014). Studies taking a longitudinal approach have found that improvement in perception preceded improvement in the production of Spanish-like voice onset time values over the course of a year (Nagle, 2018) and the same pattern was also observed in a seven-week immersion program (Casillas, 2020). However, a large amount of variability in production data is often not accounted for by the perceptual data alone. This indicates that L2 learners' production accuracy may be influenced by other factors, such as individual differences in age of arrival (Baker & Trofimovich, 2006), length of residence (Baker & Trofimovich, 2006), motor control (Kartushina et al., 2015), attitudes (Nagle, 2018), attention to acoustic cues (Huensch & Tremblay, 2015), and auditory processing abilities (Saito et al., 2022).

Methodological issues have also received increasing attention as a source of difficulty in capturing the true relationship between L2 speech perception and production (Melnik-Leroy et al., 2022; Nagle & Baese-Berk, 2022; Thomson, 2022a). For example, the discrimination (e.g., AX, ABX, oddity) and forced-choice

identification tasks most commonly used to measure perception, tap into different levels of processing, require different skills, and involve different degrees of cognitive load (Melnik-Leroy et al., 2022). Further, these perceptual measures do not have perfectly satisfactory analogs for measuring production. The tasks that are used to measure production also vary widely (e.g., elicited repetition, read aloud, picture naming, etc.). Different tasks place different cognitive demands on speakers and may not easily distinguish declarative from procedural knowledge (Llompart & Reinisch, 2019; Melnik-Leroy et al., 2022; Thomson, 2022a; Thomson & Isaacs, 2009; Thomson & Derwing, 2015).

Another methodological issue concerns the variation in evaluation of L2 speech production accuracy, ranging from acoustic measurements (e.g., formant measurements in Aliaga-García, 2017), human evaluation of linguistic features (e.g., scalar rating in Hardison, 2003), and intelligibility measured through native speaker identification of sounds (e.g., forced-choice recognition task in Bradlow et al., 1999). As Thomson (2022a) points out, the mismatch in evaluation methods for perception and production accuracy might attenuate the degree to which perception and production appear to be linked. For example, studies comparing perception measured by category discrimination with production judged by native speakers in a forced-choice identification task (e.g., Flege et al., 1999) might find a smaller perception-production correlation than would be found if perception were measured through a forced-choice identification task. Given various concerns regarding how best to evaluate perception and production and to compare the two, one of the purposes of the current meta-analysis is to explore the influence of different choices of elicitation and evaluation methods on the results of a perception-production relationship.

High variability phonetic training for improving production accuracy

As a perceptual training technique, HVPT aligns with an exemplar view of speech acquisition (Pierrehumbert, 2001; Zhang et al., 2021). Exemplar theory posits that every instance of perceptual experience including talker- and item-specific information is encoded and stored as a detailed perceptual memory (i.e., an exemplar). As a newly encountered token is compared to the exemplars already stored, a given phonetic category develops as a function of the frequency and recency of exposure to specific phonetic information. Fully developed perceptual knowledge represented by more numerous or more activated exemplars has an advantage in speech perception (in being recognized with accuracy, speed, and stability regardless of input variabilities). Given that the formation of such a robust perceptual category is a prerequisite for accurate production (Flege, 1995a; Thomson, 2022a, 2022b), HVPT can be considered a useful technique to promote production learning.

Bradlow et al. (1997) provided initial evidence that perceptual training leads to improvement in production. L1 Japanese adult learners completed a forced-choice identification training while listening to words containing an L2 English /l/ and /r/ contrast produced by five talkers in various phonetic contexts. Perception accuracy improved significantly from pretest to posttest (16.33% gain), and a relatively small but significant improvement was observed for production accuracy (5.80% gain).

Improvement in production is also generalized to untrained words. A follow-up study by Bradlow et al. (1999) further supported the efficacy of HVPT for production improvement with benefits maintained three months after training was complete. Positive evidence supporting HVPT for improvement in production, retention, and generalization has since accumulated across a variety of target sounds including vowels (Lambacher et al., 2005), stops (Carlet, 2017), fricatives (Li, 2015), nasal codas (Yang et al., 2021), lexical tones (Wang et al., 2003), and syllable structures (Huensch & Tremblay, 2015).

The degree to which improvement in perception transfers to production may vary depending on learners' individual differences and the methodology used across studies (e.g., Baker & Trofimovich, 2006; Nagle, 2018, 2021; Sakai & Moorman, 2018). For example, Bradlow et al. (1997) observed considerable variation in production gains across speakers, and the correlation between perception and production gains was nonsignificant (rho $= .202$, $p = .522$). Similarly, although not exclusively focused on HVPT studies, a meta-analysis of 18 perceptual training studies ($k = 24$) conducted by Sakai and Moorman (2018) found that perception training led to production improvement with a small effect ($d = 0.54$, see Plonsky & Oswald, 2014). In their meta-analysis, the correlation between perception and production gains at the study level ($k = 21$, $r = .31$, $p = .18$) was small and nonsignificant, although the lack of statistical significance might have been due to the relatively small sample size. Sakai and Moorman's (2018) meta-analysis further revealed that variables related to learners' individual differences and methodological choices (e.g., target phone, context of learning, explicit instruction, and learners' proficiency) significantly moderated the effectiveness of perception training in improving L2 production. Finally, we must recognize that production involves not only reference to the perceptual categories involved in perception but also articulatory component. This makes production fundamentally different, because even if individual learners' perception is accurate, the ease with which they can efficiently produce the gestures associated with individual sounds varies across sound categories and individuals (Kartushina et al., 2015).

Motivation for the present study

The current meta-analysis provides a review of studies adopting canonical HVPT defined as perception training with three central features (talker variability, phonetic context variability, and feedback) without explicit focus on production practice (Thomson, 2018). To the best of our knowledge, only two meta-analyses have previously focused on the HVPT technique (Uchihara et al., 2021; Zhang et al., 2021). Uchihara et al. (2021) meta-analyzed a total of 78 studies and found an average perception gain of 14.12% for trained stimuli and 12.96% for untrained stimuli. Their findings also confirmed the effectiveness of HVPT in terms of retention and generalization to untrained stimuli. However, their focus was exclusively on perception outcomes, not perception-to-production transfer. Zhang et al. (2021) is the only meta-analysis concerning production learning. Yet, Zhang et al.'s focus was narrow, only meta-analyzing the effect of talker variability,

comparing multi-talker and single-talker training conditions. Further, they only examined five studies with seven independent samples ($k = 7$) to find a nonsignificant and negligible effect of talker variability on production gains ($g = -0.04$; $g = -0.05$ after two outliers were removed). Moderator analysis was not conducted due to the limited sample size. Relatedly, with a relatively larger scale and a broader aim compared to Zhang et al. (2021), Sakai and Moorman's (2018) seminal meta-analysis explored how perception-based training in general promotes speech production learning. However, their wider scope of selection criteria encompassed studies that adopted a range of perception training methods beyond HVPT, including perception training with a single talker (e.g., Underbakke, 1993), listening to a story (Soler-Urzúa, 2011), and evaluating recorded speech of other students for nativelikeness (Counselman, 2010). Given that HVPT is determined by the three core components of: (a) variability in talkers, (b) variability in phonetic context, and (c) immediate feedback (Thomson, 2018, 2022b), the results of their meta-analysis are unlikely to provide direct insights into the effectiveness of HVPT in improving L2 production (see Appendix 1 for the summary table of studies included in Sakai and Moorman's synthesis).

Thus, the current meta-analysis focusing exclusively on HVPT is unique and methodologically distinctive from Sakai and Moorman (2018) in that the three core features are held constant. Notably, our attempt responds to the call for the application of HVPT in pedagogical practice (Barriuso & Hayes-Harb, 2018; Thomson, 2018) and the sharp increase in the number of studies adopting the HVPT procedure (12 HVPT studies in Sakai and Moorman, 2018 vs. 31 HVPT studies in the current meta-analysis). Utilizing a larger sample size ($k = 43$) of studies from the past 32 years (1991 to 2022), our primary goal of this meta-analysis is to establish the overall effectiveness of HVPT for production learning (production gains, retention, and generalization). We also aim to clarify learner-related and methodological variables that may promote (or hinder) the beneficial effects of HVPT. As such, this meta-analysis is designed to determine (a) whether HVPT leads to improvements in production; if so, to what extent features related to (b) learner, (c) training, and (d) outcome measures modulate the amount of improvement. This research is expected to provide pedagogical guidance on best practices of HVPT for improving learners' speech production.

Accordingly, the current study is guided by the following research questions (RQs):

1. How effective is HVPT without explicit focus on production practice for improving production accuracy of L2 sounds?
2. Is the improvement in L2 production accuracy retained over time?
3. Is the improvement in L2 production accuracy generalized to new stimuli?
4. Is there a relationship between perception and production gains?
5. What features related to learner, training, and outcome measures of HVPT influence production gains?

Method

Replication package

Completed coding sheets and descriptive statistics for each coded study (e.g., effect sizes, standard errors, and variances) are accessible at https://osf.io/4fu3c/. Analysis codes for effect-size calculation and moderator analyses from the software used in this study (Comprehensive Meta-Analysis Software, https://www.meta-analysis.com/) were not accessible. Thus, we have provided a detailed description of equations used for calculating effect sizes and step-by-step procedures for conducting all statistical analyses using the software with raw data files (.cma) for replication purposes. See Supplementary Materials for a summary of equations used to calculate effect sizes (Appendix 2) and a list of 31 studies included in the meta-analysis (Appendix 3).

Literature search

When searching for literature, all manuscript types were considered for this meta-analysis (i.e., journal articles, book chapters, conference proceedings, and graduate dissertations). The search process consisted of searching prominent databases, reviewing previous syntheses on HVPT, and searching/reviewing table of contents of prominent journals (see Figure 1 for full details). Keywords for searches included: HVPT, high variability phonetic training, high variability perceptual training, high variability segmental training, L2 speech perception training, and computer-assisted pronunciation training. Searches were conducted by the first two authors and constrained to between 1991, the year of Logan et al.'s (1991) seminal study, and December 2022. To ensure no manuscripts were missed, final searches were conducted on Google and Google Scholar.

Inclusion and exclusion criteria

We identified 1,675 reports as potentially eligible to be included in the meta-analysis. The first and second authors of this study screened the studies with the following inclusion and exclusion criteria, resulting in a total of 31 reports available for the current meta-analysis:

(a) The study was an empirical investigation of canonical or perceptual HVPT, defined as perceptual training in which listeners received auditory stimuli produced by two or more talkers in multiple phonetic environments. Studies involving auditory training stimuli produced by a single talker were excluded. Studies that included production training in the training program that could not be separated from the perception training were excluded (e.g., Mora et al., 2022). This decision was made primarily because of the concern about the role of self-monitoring production as an additional input and the potentially negative impact of such input on production learning (Thomson, 2022b).

(b) The study reported production accuracy. Studies that reported only perception accuracy data were excluded.
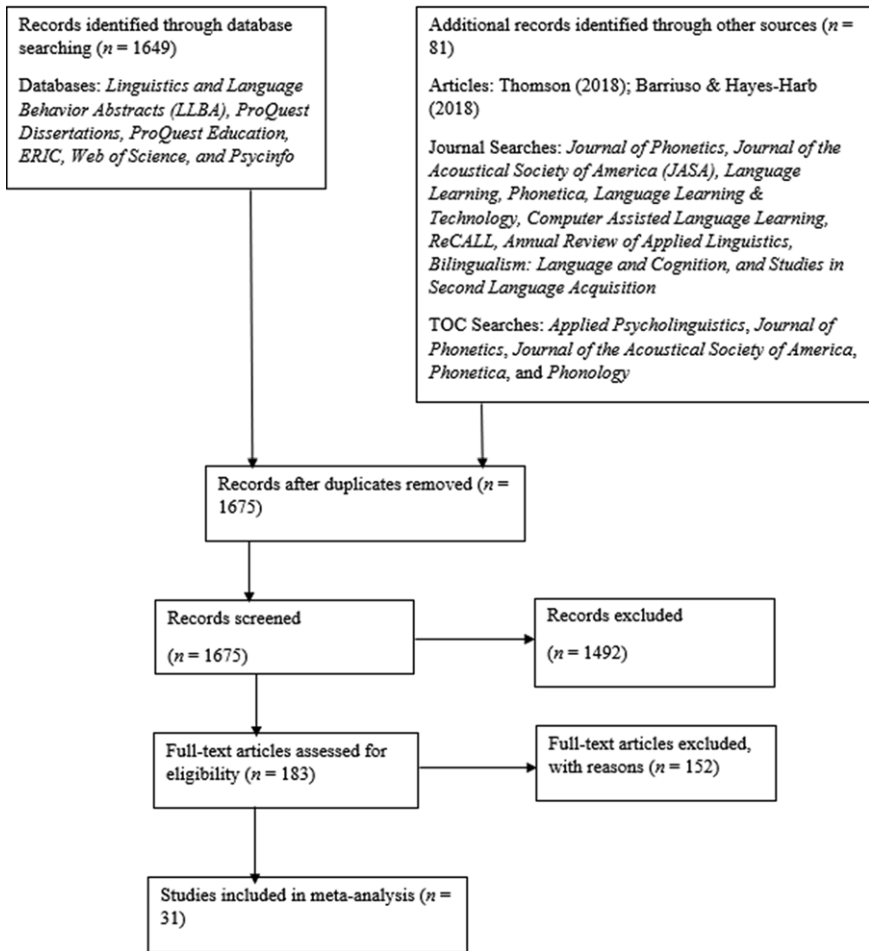
Search Process

Records identified through database searching (*n* = 1649)	Additional records identified through other sources (*n* = 81)
Databases: *Linguistics and Language Behavior Abstracts (LLBA), ProQuest Dissertations, ProQuest Education, ERIC, Web of Science, and Psycinfo*	Articles: Thomson (2018); Barriuso & Hayes-Harb (2018) Journal Searches: *Journal of Phonetics, Journal of the Acoustical Society of America (JASA), Language Learning, Phonetica, Language Learning & Technology, Computer Assisted Language Learning, ReCALL, Annual Review of Applied Linguistics, Bilingualism: Language and Cognition, and Studies in Second Language Acquisition* TOC Searches: *Applied Psycholinguistics, Journal of Phonetics, Journal of the Acoustical Society of America, Phonetica,* and *Phonology*

Records after duplicates removed (*n* = 1675)

Records screened (*n* = 1675)	Records excluded (*n* = 1492)

Full-text articles assessed for eligibility (*n* = 183)	Full-text articles excluded, with reasons (*n* = 152)

Studies included in meta-analysis (*n* = 31)

Figure 1. Literature search procedure.

    (c) The study involved a within-participant design (i.e., pretest-posttest contrast) and/or a between-participant design (i.e., treatment-control contrast). The control group was defined as the group of participants taking perception tests without completing perceptual training or completing perceptual training focusing on untargeted sounds.[1]
- When the data for a between-participant design were available, the study reporting gain scores (i.e., pretest-posttest mean differences) for both treatment and control groups was included. Studies comparing only post-training performance between the two groups were excluded.

    (d) Studies providing phonetic information (e.g., providing a brief description of articulatory gestures to produce target sounds before training, Rato, 2014) or

providing audiovisual information during training (e.g., seeing a talker's face, Hazan et al., 2005) were included.[2]

(e) The study provided trial-by-trial feedback. Studies that did not provide feedback during training were excluded.

(f) The study employed behavioral tasks that reported production accuracy. We adopted a comprehensive approach to include studies using various types of production elicitation tasks (e.g., word reading, sentence reading) and outcome measures (e.g., native speaker identification, acoustic analysis).

(g) The study focused on the learning of segmentals (vowels and consonants) and/or suprasegmentals (lexical tones and syllable structures). Studies that focused on linguistic aspects beyond segmental and suprasegmental features (e.g., dialect, vocabulary) were excluded.

(h) The study involved participants without any reported language or hearing disabilities whose L2 (not L1) perception was trained.

(i) The report had to be written in English.

(j) The study had to report sufficient statistical information (e.g., sample size, means, standard deviations, t value, and F value) to calculate effect sizes.

Coding

The coding scheme for this study was first created based on previous systematic reviews of HVPT and perception training (Barriuso & Hayes-Harb, 2018; Sakai & Moorman, 2018; Thomson, 2018) and modified through an iterative process of coding and revision. From the included studies, data were extracted and coded for study identifiers (authors, year, and publication type) and moderator variables.

Learner features

Regarding learner-related variables, age of testing (AOT), learning context, and proficiency were coded. AOT was defined as the age at which participants completed perception training. Learning context was coded as to whether participants resided in foreign language (FL or a country where L2 is not spoken) or second language (SL or a country where L2 is spoken) contexts at the time of testing. Proficiency was first categorized into six levels (novice vs. beginner vs. lower intermediate vs. intermediate vs. upper intermediate vs. advanced). Because the number of samples for some of the categories was limited (k for novice and upper intermediate = 2 respectively), these categories were combined to re-create three proficiency levels: lower level (novice + beginner + lower intermediate), intermediate, and higher level (upper intermediate + advanced). Studies involving learners with different proficiency levels were coded as mixed.

According to the hypothesis of Iverson et al. (2012), HVPT may have little effect on experienced learners residing in the L2 naturalistic environment because it provides the opportunity for learners to receive a wider range of phonetic input in daily life. It was therefore expected that higher-proficiency learners living in SL contexts would benefit less from HVPT compared to lower-proficiency learners in FL contexts (but see Georgiou, 2021 for counterevidence in the case of L2 vowel

discrimination). Regarding AOT, given the limited age ranges of the participants in this synthesis (age = 12 to 37), a strong effect of AOT was not expected. However, based on the SLM (Flege, 1995a), younger learners were predicted to benefit more from HVPT than older learners (e.g., see Shinohara & Iverson, 2021 for evidence showing the advantage for adolescents [15-18 years of age] over adults [25 years of age] in perception gains).

Training features

Motivated by Thomson's (2018) critical review of HVPT studies, we selected seven variables related to training features (type of test items, target phones, phonetic information, training duration, number of talkers, training time per session [min], and total training time [min]). The goal of this analysis is practically oriented in order to examine ways to optimize the efficacy of HVPT for production learning. The type of test items were coded as either old or new items (in response to RQ3 regarding generalization of learning). New items were defined as stimuli that did not occur during training, whereas old items were stimuli that occurred during training. Target phones were coded for segmental (obstruent, sonorant, and vowel) and suprasegmental features (lexical tone and constraint on syllable structure). Phonetic information was coded dichotomously regarding whether participants were provided with articulatory information about target sounds before training. Training duration was coded categorically (the training lasted for less than 1 month vs. 1 month or longer).

Production measures

One important responsibility of meta-analysts is scrutinizing the methodological variabilities across studies with the aim of providing methodological guidance for future research. Thus, we followed the common practice of earlier meta-analyses on L2 pronunciation learning (Lee et al., 2015; Sakai & Moorman, 2018; Saito & Plonsky, 2019; Saito, 2021) to focus on examining the effects of variables related to the choice of production measures.

We examined four variables related to production measures (outcome measures, rater experience, elicitation tasks, and prompt modality). Outcome measures were coded as acoustic analysis, native speaker identification, scalar rating, or transcription of L2 speech. Production elicitation tasks included word reading, sentence reading, recalling, and repetition. Prompt modality concerned whether written or spoken input (or both) was presented in eliciting the production of target sounds.

Rater experience was defined as whether raters received training, had language teaching experience, or education in phonetics. Raters who did not receive any training or had no experience in L2 teaching or phonetics were classified as inexperienced raters. The L2 speech literature has suggested that experienced raters who are familiar with foreign accents (Kennedy & Trofimovich, 2008) and have a linguistics and/or teaching background (Saito et al., 2017) tend to be more lenient in evaluating L2 speech performance. Despite the varying degrees of leniency depending on rater backgrounds, untrained or inexperienced raters generally provide consistent and reliable judgments (Derwing et al., 2004; but see Isaacs &

Thomson, 2013 for evidence pointing to a qualitatively different rating process). Given that the sufficient rating consistency can be expected from inexperienced raters, we predicted that there would not be any major impact on rater experience.

The first two authors coded the studies independently in the first round. Intercoder agreement of 100% was achieved for all variables except L2 proficiency. To ensure the accuracy of coding for proficiency, a second round of coding was completed. The intercoder reliability between the two coders was 96.78% with the sole disagreement discussed and agreed upon.

Data analysis

To compute the weighted mean effect size and conduct moderator analysis, we employed the Comprehensive Meta-Analysis (Version 3.3) software. We relied on Hedges' g as the basic unit of analysis, a transformed version of Cohen's d which corrects for bias in small samples. To answer RQ1 regarding the transfer of HVPT to L2 production, we conducted two separate analyses according to different study designs: within-participant effect sizes (i.e., the mean difference between pretest and posttest scores) and between-participant effect sizes (i.e., the mean difference in gain scores between HVPT treatment and control groups). Several studies reported multiple descriptive data from the same participants (the use of multiple elicitation tasks such as word reading and sentence reading, outcome measures such as identification and transcription, prompt modality, target phones, and test item types). Thus, such multiple scores except outcome measures were averaged to yield a single score to avoid violating the requirement of independence of observations. When multiple outcome measures were available, we followed Sakai and Moorman (2018) to select the elicited production with native speaker identifications as the representative task. As a result, a total of 43 independent effect sizes were available for the analysis of within-participant data, and 17 independent effect sizes were available for the analysis of between-participant data. The L2 field-specific benchmarks for independent standardized mean differences were used to interpret the findings (.40 for small, .70 for medium, and 1.00 for large effects; Plonsky & Oswald, 2014).

To answer RQ2 regarding the retention of gains in production, we computed the weighted standardized mean difference (a) between post-training accuracy scores and delayed posttest scores and (b) between pretest scores and delayed posttest scores. Studies reporting delayed posttest scores were first selected for this analysis ($k = 10$). The same procedure as used for calculating the pretest-posttest mean difference to answer RQ1 was adopted to calculate the two types of effect size (posttests vs. delayed posttests and pretests vs. delayed posttests). The mean interval between posttests and delayed posttests was 1.4 months (range = 0.5 to 3 months).

To answer RQ3 (examining the generalization of production learning) and RQ5 (examining features of HVPT that predict production gains), we used a mixed-effects model to conduct moderator analysis with 10 categorical and four continuous variables. Moderator analyses were conducted with a between-group Q statistic for predetermined categorical variables. For continuous variables, meta-regression analyses were conducted with a full Maximum Likelihood method. The moderator analyses were conducted only for the within-participant data given the

statistical robustness with the larger sample size ($k = 43$) compared to the between-participant data ($k = 17$). In the moderator analysis of outcome measures, production elicitation tasks, prompt modality, and test item types, we followed Sakai and Moorman (2018) to include a separate analysis of effect sizes from all studies reporting the information about the four variables without averaging them to yield a single score per study.

To answer RQ4 regarding the relationship between perception and production gains, we adopted two levels of analysis (i.e., participant and study level). The first approach at the participant level was to focus on 21 studies reporting correlation coefficients between perception and production gains. With the Comprehensive Meta-Analysis (Version 3.3), we conducted a random-effects model to compute the weighted mean correlation. In order to provide a fuller picture of the extent to which perception and production accuracy are associated at different time points of testing, the same analyses were conducted for the studies reporting the correlation for pretest performance ($k = 14$) and for posttest performance ($k = 13$). The second approach was to focus on a study-level correlation between perception and production gain scores (see Sakai & Moorman, 2018). In 38 studies, the standardized mean difference between pretest and posttest scores (or within-participant Hedges' *g*) was available for both perception and production data. We conducted a Pearson correlation analysis between perception and production effect sizes for post-training data ($k = 38$) and a Spearman correlation analysis for retention data ($k = 14$). The effect size was interpreted based on Plonsky and Oswald's (2014) benchmarks: small ($r = .25$), medium ($r = .40$), and large ($r = .60$).

Results

Description of included studies

Thirty-one articles published between 1991 and 2022 passed all inclusion and exclusion criteria with the exception of reporting sufficient data for the quantitative synthesis. These 31 reports were published journal articles ($n = 16$), doctoral dissertations ($n = 10$), conference proceedings ($n = 4$), and a book chapter ($n = 1$). The average sample size per study was 17.06 ($SD = 9.76$). Twenty-nine out of 31 articles reported production accuracy scores by native speaker identification in percentage, out of which the information of test item types (old and new items) was available. The unweighted mean gain from pretests to posttests was 10.56% ($k = 9$, $SD = 8.60\%$, 95% CI [3.94, 17.17]) for old items and 4.5% ($k = 12$, $SD = 3.94\%$, 95% CI [1.99, 7.01]) for new items. Production gains appeared to be higher when learners produced trained test items than when producing untrained test items. Thirty-one reports yielded 43 independent effect sizes for pretest-posttest gains, of which 17 studies compared production gains from treatment and control groups.

Forty-three experiments targeted participants with various L1 backgrounds, including Japanese ($k = 10$), Korean ($k = 7$), English ($k = 7$), Chinese ($k = 6$), Catalan-Spanish ($k = 5$), Spanish ($k = 4$), and Greek and Thai ($k = 4$). The vast majority focused on L2 English ($k = 33$), followed by Mandarin Chinese ($k = 5$), Japanese ($k = 3$), Spanish ($k = 1$), and French ($k = 1$). Of the studies that reported

definitive numbers, the average age at which learners participated in the training experiments was 23.89 years old (range = 12–37). About half of the experiments focused on university students ($k = 26$), and other experiments targeted students in secondary school ($k = 4$), primary school ($k = 1$), and language institutes ($k = 7$).

Regarding the materials and procedures for perception training, all experiments adopted identification training tasks except for two which utilized discrimination training tasks. The experiments focused on vowels ($k = 20$), obstruents ($k = 8$), sonorants ($k = 8$), lexical tones ($k = 3$), and syllable structures ($k = 2$). The number of talkers ranged from 3 to 30 ($M = 6.6$) with the majority of experiments involving 4 talkers ($k = 21$). Training intensity varied across experiments. The average number of training sessions was 11 ($SD = 8.7$; range = 3–45 sessions), the average amount of training time per session was 35.77 min ($SD = 17.7$; range = 10–75 min), and the average total amount of training time was 310.64 min ($SD = 204.8$; range = 60–1125 min).

With respect to production elicitation and outcome measures, all experiments utilized controlled production tasks such as word reading ($k = 21$), sentence reading ($k = 11$), repetition ($k = 7$), and recalling ($k = 5$). All repetition tasks except Bradlow et al. (1999) were considered delayed repetition as efforts were made to avoid the influence of auditory input of a native speaker model by, for example, inserting a 3000 ms interval (Aliaga-García, 2017) or a white noise distractor (Dong et al., 2019) before pronouncing target items. Elicited speech samples were evaluated by means of native speaker identification ($k = 23$), scalar rating ($k = 21$), orthographic or phonetic transcription ($k = 4$), and acoustic analysis ($k = 3$). Experiments using native speaker identification involved an average of 20.2 raters (range = 2–110) and the average number of raters per speech sample was 7 (range = 1–28). Interrater reliability was reported in 10 out of 23 experiments (43%) using a wide range of measures (i.e., Cronbach's alpha, Kendall's W, Cohen's kappa, correlation coefficient, and percentage agreement). Experiments employing scalar rating assessment adopted different number of scale points ($M = 6.5$, range = 3–11) with production accuracy defined as comprehensibility (e.g., how easy or difficult to identify L2 sounds: Carlet, 2017), nativelikeness (e.g., native-like vs. clearly not native: Macdonald, 2012), or simple correctness (e.g., bad vs. excellent: Hazan et al., 2005). The average number of raters per experiment was 14.6 (range = 2–40), and the average number of raters per speech sample was 7.4 (range = 2–26). Interrater reliability was reported in 13 out of 21 experiments (62%) using intraclass correlation coefficients, Cronbach's alpha (range = 0.78–0.85), and percentage agreement. Four experiments had raters (range = 3–110 per experiment and range = 3–20 per speech sample) orthographically or phonetically transcribe elicited samples, out of which two experiments from one study reported interrater reliability (i.e., Cronbach's alpha).

RQ1: How effective is HVPT for improvement of production accuracy?

To examine the transfer of HVPT to production, we conducted two sets of effect-size aggregation analyses, one for within-participant effect sizes and one for between-participant effect sizes. First, we inspected 43 effect sizes of production gain scores for potential outliers. Li (2015), considered a potential outlier in the current
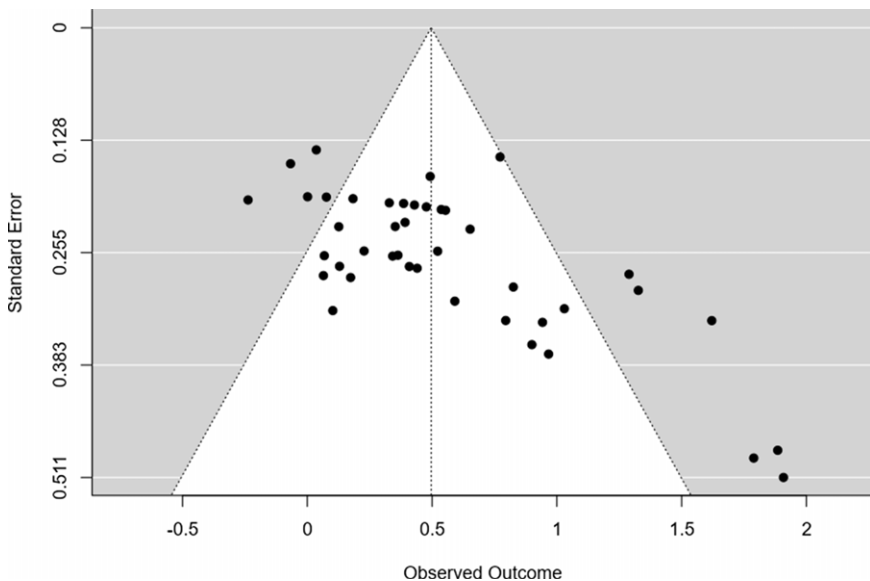
Figure 2. Funnel plot of production effect sizes (mean pretest-posttest differences) by standard error.

data ($> 3SD$), would have a misleadingly substantial impact on the overall mean effect of HVPT ($g = 0.49 \rightarrow 0.56$). Because the overestimation of the mean effect due to the single study sample was concerning, Li (2015) was identified as an outlier and removed from the subsequent analysis. A follow-up inspection was conducted by repeatedly calculating the mean effect size after removing every study one at a time to gauge the impact of a single effect size on the overall mean result. The one-study-removed analysis showed no substantial change in the mean effect size in the remaining 42 samples (range $= 0.47$ to 0.51), indicating no further obvious outliers in the final data set. The result based on the 42 effect sizes showed a significant and small effect of training on production improvement, $g = 0.49$, $SE = 0.07$, 95% CI [0.36, 0.62], $p < .001$. Publication bias was assessed with a funnel plot (Figure 2). Three studies with a lower precision (or higher SE) appear to produce larger effect sizes. The Egger's test was significant, suggesting the presence of publication bias, $t(40) = 4.66$, $p < .001$. However, the trim-and-fill method identified no studies to be imputed, and the classic fail-safe N test showed that 1,726 studies would be needed to nullify the significant training effect at $p > .050$. These results indicate that the issue with publication bias was not considered serious in our data set.

In addition to estimating the mean effect size for the pretest-posttest difference, we aggregated the 17 independent effect sizes for the difference of production gain scores between the treatment and control groups. No studies were identified as outliers, and the one-study-removed analysis showed no substantial change in the data set of 17 observed effect sizes (range $= 0.58$ to 0.71). The result of effect-size aggregation showed a significant and medium effect of training on production improvement, $g = 0.66$, $SE = 0.12$, 95% CI [0.42, 0.90], $p < .001$, indicating that treatment groups receiving HVPT improved to a greater degree compared to
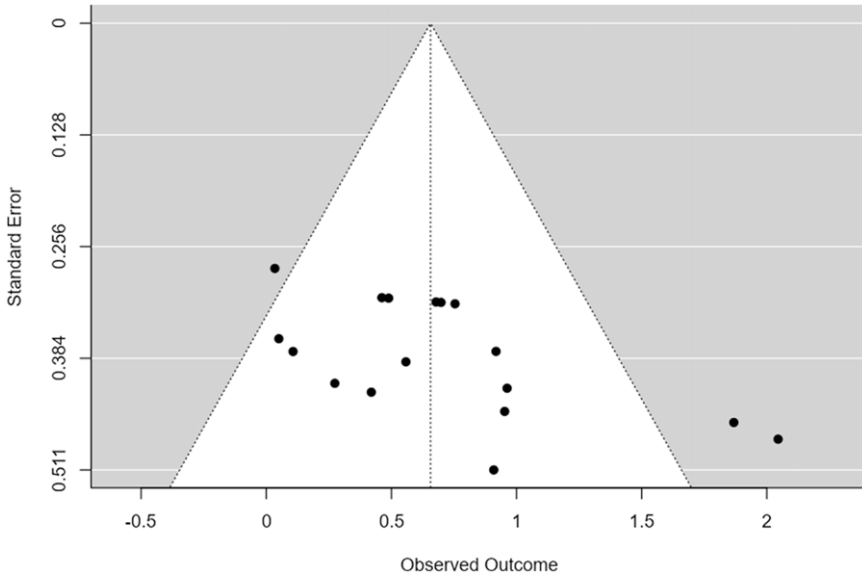
Figure 3. Funnel plot of production effect sizes (mean treatment-control differences) by standard error.

control groups. Publication bias was assessed with a funnel plot (Figure 3). The Egger's test was significant, $t(15) = 2.78$, $p = .014$, indicating the presence of publication bias. However, the trim-and-fill method identified no studies to be imputed, and the classic fail-safe N test showed that 224 studies would be needed to nullify the significant training effect at $p > .050$. These results indicate that the issue with publication bias was not considered serious in our data set.

RQ2: Is the production gain retained?

Based on 10 studies reporting production scores at pretest, posttest, and delayed posttest, two analyses of effect-size aggregation were conducted, one for pretest-delayed posttest comparison and one for posttest-delayed posttest comparison (mean retention interval = 1.4 months, range = 0.5 to 3 months). The mean effect size for the production gain from pretest to delayed posttest was small and approached statistical significance, $g = 0.26$, $SE = 0.13$, 95% CI [0.00, 0.52], $p = .053$. The mean effect size for the contrast of posttest and delayed posttest performance was negligible and not significant, $g = -0.10$, $SE = 0.15$, 95% CI [−0.39, 0.19], $p = .506$. In summary, the result that the post-training effect initially observed in answer to RQ1 ($g = 0.49$) decayed approximately after 1.4 months ($g = 0.26$), despite the lack of significant difference between posttest and delayed posttest performance, does not provide strong support for the retention of production learning.

RQ3: Is the production gain generalized?

This analysis was based on 40 within-participant studies reporting the test stimuli type (old or new items). The mean effect size was calculated for old items (i.e., items that occurred during training, $k = 26$) and new items (i.e., items that did not occur during training, $k = 14$) separately. The mean effect size for old items was medium and significant, $g = 0.61$, $SE = 0.09$, 95% CI [0.43, 0.79], $p < .001$, whereas for new items the training effect was small and approached significance, $g = 0.20$, $SE = 0.12$, 95% CI [−0.03, 0.43], $p = .091$. The difference in the training effects between old and new items was significant, $Q(1) = 7.76$, $p = .005$, indicating that production of L2 sounds was more accurate when learners produced trained items than when producing untrained items.

RQ4: Is there a relationship between perception and production gains?

The relationship between perception and production gains was examined at the participant and study levels. At the participant level, 21 studies reported the correlation between perception and production gains. The effect-size aggregation showed that the mean correlation was negligible and nonsignificant, $r = .09$, 95% CI [−.02, .20], $p = .122$. The heterogeneity test showed that the variation in effect size was negligible, $Q(20) = 14.86$, I-squared $= 0.00$, $p = .784$. To explore an overall picture of the perception-production connection at different times of testing, two additional analyses focusing on the pretest- and posttest-level correlation respectively were also conducted. The mean effect size was significant and small to medium for the pretest correlation ($k = 14$, $r = .41$, 95% CI [.23, .56], $p < .001$) and for the posttest correlation ($k = 13$, $r = .33$, 95% CI [.19, .46], $p < .001$).

At the study level, we conducted correlational analysis between the effect sizes (Hedges' g) of perception and production improvement (pretest-posttest improvement: $k = 38$) and retention (pretest-delayed-posttest improvement: $k = 14$). Two studies (Li, 2015; Reyes & Hazan, 2021) were identified as outliers (> 3 SD) and removed from the analysis ($k = 36$). Figure 4 shows a linear relationship between perception improvement (unweighted mean $= 1.08$, 95% CI [0.89, 1.26]) and production improvement (unweighted mean $= 0.56$, 95% CI [0.40, 0.72]), indicating that studies which show a larger improvement in perception accuracy were more likely to demonstrate a larger improvement in production accuracy. This trend was confirmed with the result of Pearson correlation analysis showing a significant and medium effect size ($r = .45$, 95% CI [.14, .68], $p = .006$). Regarding the retention data (see Figure 5), Spearman's correlation analysis was conducted due to small sample size ($k = 14$), showing a significant and large correlation between perception and production retention (rho $= .78$, 95% CI [.42, .93], $p < .001$).

RQ5: Which features of HVPT influence production gains?

Moderator analysis was conducted on nine categorical variables. The results are summarized in Table 1. The effect of learning context was approaching significance ($p = .082$), with a larger effect of training observed for learners in FL contexts ($g = 0.60$) compared to those in SL contexts (g $= 0.34$). L2 proficiency and target phones were not significant moderators of production gains ($p > .050$), indicating
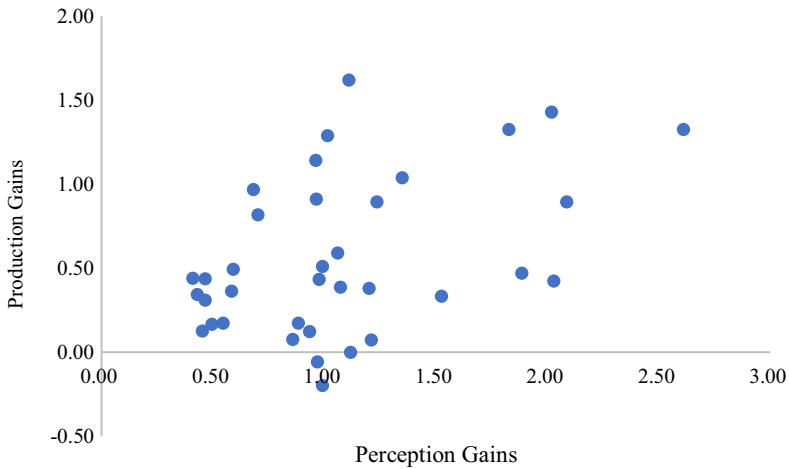
Figure 4. Scatterplot of perception and production gains (pretest-posttest improvement).
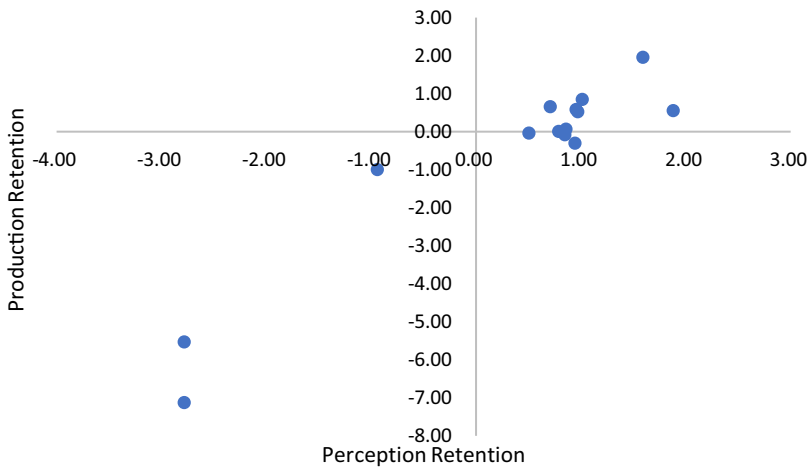

Figure 5. Scatterplot of perception and production retention (pretest-delayed-posttest improvement).

that HVPT appears to work uniformly regardless of learners' proficiency levels or target L2 sounds. Provision of phonetic information was a significant moderator ($p = .029$), indicating that a training effect was larger when phonetic information was provided ($g = 0.94$) than when it was not provided ($g = 0.45$). Regarding training duration, a larger effect was found when training lasted for less than one month ($g = 0.61$) compared to when it continued for one month or longer ($g = 0.32$), and the difference approached significance ($p = .054$). The use of different outcome measures significantly moderated production gains ($p = .012$), with transcription producing a larger effect ($g = 1.14$) compared to acoustic analysis ($g = 0.18$), identification ($g = 0.45$), and rating ($g = 0.28$). No significant effect of rater experience was found, indicating that results were not greatly

Table 1. Moderator analyses for categorical variables (pretest-posttest comparison)

Variables	k	g	SE	95% CI LL, UL	p	Q tests Q	p
Learning context						3.02	.082
FL	27	0.60	0.08	0.43, 0.76	<.001		
SL	12	0.34	0.12	0.10, 0.58	.006		
Proficiency						1.34	.719
Higher level	11	0.55	0.13	0.30, 0.81	<.001		
Intermediate	6	0.51	0.18	0.15, 0.87	.005		
Lower level	11	0.45	0.13	0.21, 0.70	<.001		
Mixed	8	0.34	0.14	0.06, 0.62	.016		
Target phone						0.39	.983
Obstruent	5	0.43	0.20	0.03, 0.83	.034		
Sonorant	8	0.55	0.17	0.22, 0.88	.001		
Syllable	2	0.65	0.33	0.01, 1.30	.048		
Tone	3	0.55	0.27	0.02, 1.08	.043		
Vowel	20	0.53	0.10	0.33, 0.74	<.001		
Phonetic information						4.47	.029
Yes	4	0.94	0.21	0.52, 1.36	<.001		
No	38	0.45	0.07	0.32, 0.58	<.001		
Duration						3.71	.054
≥ 1 month	12	0.32	0.12	0.09, 0.55	.006		
< 1 month	24	0.61	0.09	0.43, 0.79	<.001		
Outcome measure						11.04	.012
Acoustic analysis	3	0.18	0.24	−0.29, 0.66	.438		
Identification	23	0.45	0.09	0.29, 0.63	<.001		
Rating	20	0.28	0.09	0.10, 0.46	.002		
Transcription	4	1.14	0.26	0.64, 1.65	<.001		
Rater experience						0.32	.574
Experienced	24	0.44	0.09	0.26, 0.62	<.001		
Inexperienced	13	0.36	0.12	0.12, 0.59	.003		
Elicitation task						5.43	.143
Recall	5	0.58	0.20	0.18, 0.97	.005		
Repetition	7	0.28	0.17	−0.06, 0.62	.105		
Sentence reading	10	0.45	0.15	0.17, 0.74	.002		
Word reading	21	0.72	0.11	0.51, 0.94	<.001		

(*Continued*)

Table 1. (*Continued*)

Variables	k	g	SE	95% CI LL, UL	p	Q tests Q	p
Prompt modality						4.89	.087
Auditory input	6	0.19	0.17	−0.15, 0.54	.265		
Written input	35	0.62	0.08	0.46, 0.77	.012		
Auditory + written input	4	0.56	0.22	0.12, 1.00	.012		

Notes: CI = confidence interval; *SE* = standard error; LL = lower limit; UL = upper limit.

influenced by whether L2 sounds were evaluated by experienced ($g = 0.44$) or inexperienced raters ($g = 0.36$). Although elicitation tasks did not exert a significant impact on production gains ($p = .143$), word reading yielded a relatively larger training effect ($g = 0.72$) compared to recalling ($g = 0.58$), repetition ($g = 0.28$), and sentence reading ($g = 0.45$). Further analysis of prompt modality in eliciting the production showed that input modality was a marginally significant moderator of production gains ($p = .087$). The training effect was smaller when L2 sound production was elicited only through auditory input ($g = 0.19$) compared to only through written input ($g = 0.62$) or combined written and auditory input ($g = 0.56$).

Second, meta-regression analysis was conducted to examine whether four continuous variables predict production gains. The age at which learners participated in perception training (AOT) was negatively associated with production gains with marginal significance, $k = 38$, coefficient $= -0.024$, $SE = 0.013$, 95% CI [−0.049, 0.000], $p = .053$, indicating that younger participants benefited from HVPT to a greater extent than older participants. The number of talkers was not a significant predictor of production gains, $k = 40$, coefficient $= -0.012$, $SE = 0.011$, 95% CI [−0.033, 0.009], $p = .250$. Although no significant impact of total training time was found, $k = 38$, coefficient $= 0.000$, $SE = 0.000$, 95% CI [−0.001, 0.001]. $p = .524$, training time per session was negatively associated with production gains, $k = 38$, coefficient $= -0.009$, $SE = 0.004$, 95% CI [−0.016, −0.001], $p = .022$. These results indicate that production gains tended to become smaller when learners spent more time on perception practice in each session, while the total amount of time they spend on practice did not promote or hinder the gains in production.

Discussion

How effective is HVPT for improving speech production accuracy?

In answer to RQ1, small-to-medium effects of HVPT for production gains were found ($g = 0.49$ for pretest-posttest comparison, $g = 0.66$ for treatment and control comparison). These findings are consistent with the previous meta-analysis by Sakai and Moorman (2018) focusing on studies of general perception training ($d = 0.54$ for pretest-posttest comparison; $d = 0.89$ for treatment and control comparison). The relatively smaller gains for production were observed compared

to perception ($g = 0.96$ for pretest-posttest comparison; $g = 0.97$ for treatment and control comparison) according to a recent meta-analysis of HVPT for perception learning (Uchihara et al., 2021).

In answer to RQ2 and RQ3 regarding retention and generalization, the analyses indicate that the effect of perception-to-production transfer may not be long-lasting, and production gains may not be generalizable to untrained stimuli. Although no clear difference was found between posttest and delayed posttest accuracy ($g = -0.10$), the long-term improvement from pretest to delayed posttest was considered very small ($g = 0.26$) and appeared to be on a trajectory of regression towards the learners' starting point prior to training.

As for generalization to new phonetic contexts/words, the medium effect ($g = 0.61$) was observed for learning old items (i.e., stimuli that appeared during training), whereas a marginally significant and small effect ($g = 0.20, p = .091$) was found for learning new items (i.e., stimuli that did not occur during training). These findings contrast with previous meta-analysis findings for perception learning (Uchihara et al., 2021) that found, for retention, perception at the delayed posttest was much more accurate than perception at the pretest ($g = 0.98$) and a negligible difference was observed between posttest and delayed posttest ($g = -0.08$, $p = .058$). For generalization, perception gains for old items and old talkers ($g = 0.91$) and for new items and new talkers ($g = 0.96$) were almost comparable.

In answer to RQ4 regarding the correlational link between perception and production gains, the results of correlation analyses between perception and production were not consistent across levels of analysis. At the participant level, a nonsignificant and negligible relationship was found ($r = .09$). At the study level, a significant and medium-sized relationship ($r = .45$) was found for perception and production gains (comparable to the finding of Sakai & Moorman, 2018: rho $= .39$), and a larger correlation was observed for retention data (rho $= .78$). Although the relatively large correlations observed at the study level seem to suggest the strong link between speech perception and production, the results need to be interpreted with caution. The study-level results may simply indicate that the studies with the higher quality of training setup and implementation are likely to improve both perception and production accuracy (Thomson, 2012b), not necessarily suggesting a causal relationship between perception and production.

The findings can be summarized as follows: (a) perception-only training led to moderate production gains; (b) strong support for the retention of production gains was lacking; (c) production learning was not generalized to new items; (d) the correlation between perception and production gains at the participant level was negligible, although moderate-to-large correlations were found at the study level. These findings collectively do not provide strong evidence for perception-to-production transfer as a result of exposure to high variability input with corrective feedback.

The failure of transfer may be largely because of the time-lagged nature of the relationship between perception and production development (Nagle, 2018, 2021; Nagle & Baese-Berk, 2022). According to the SLM (Flege, 1995a) and SLM-r (Flege & Bohn, 2021), the mechanisms underlying L1 speech learning are still available to L2 learners, but our findings suggest that accessing production mechanisms, while still possible, is more difficult than accessing perceptual mechanisms. This is likely because the two mechanisms are distinct despite being connected. Perceptual

changes are cognitive, while articulatory processes include both cognitive and physical components. On the articulation side, motor movements in a learner's L1 are so automatic that initiating a new articulatory command is difficult after a certain age (Scovel, 1969). This is true of motor skills that are unrelated to speech, such as athletes attempting to change an established running gait or golfers modifying their swings, etc. Relatedly, the time lag between improvement in L2 perception and production may be accentuated to a greater extent, given that learners' own productions may become the primary input to their new L2 system. L2 listeners without articulatory control fine-tuned to the L2 system may bring their own perceptual distortions to the production task (Thomson, 2022a). Conversely, the current findings did not provide strong support for the gesturalist claim that perception of L2 speech relies on knowing how to produce the sound using articulatory gestures (PAM-L2: Best & Tyler, 2007). If perception learning required the perceiver to fully understand how to use the articulatory gestures, a stronger perception-to-production transfer in terms of immediate gains, retention, and generalization would have been observed.

Alternatively, the correlational data at the participant level suggest a stronger link between speech perception and production at a fixed point in time. Despite the absence of a significant correlation for gain scores, significant and moderate overall correlations were found for pretest ($r = 0.41$) and posttest ($r = 0.33$) data. Because these correlations do not reflect perception and production gains as a result of training, the findings indicate that learners who can accurately perceive target sounds tend to produce the sounds accurately at a fixed point in time. These findings, especially the fact that the correlation at the pretest is slightly higher than that at the posttest, imply that the perception-production link may emerge on the robust and stable representations developed through past L2 experience for a long term. In order to see a stronger link in training data, a rigorous longitudinal study tracking the progress in perception and production over an extended period of time is warranted (e.g., Nagle, 2018).

What factors moderate the transfer of HVPT to L2 production?

To answer RQ5, we explored the effects of nine categorical and four continuous variables on production improvement. L2 proficiency and target phones were not found as significant variables, indicating that HVPT seems to be beneficial in production learning for learners with different proficiency levels and for various target sounds including segmental (i.e., vowels and consonants) and suprasegmental features (i.e., lexical tones and syllable structures). This aligns with an exemplar view of speech learning, or Flege (1995a)'s emphasis on the role of relative experience with oral language. While proficiency in English varies across learners, this is often rooted in differences in written language proficiency (especially in FL contexts), which will not provide the sort of auditory input necessary for the development of L2 speech categories. As such, an L2 learner may develop advanced proficiency in reading, vocabulary, and grammar without much experience in listening. Such advanced proficiency learners appear to benefit from HVPT as much as lower-proficiency learners.

Rater experience did not significantly affect perceived intelligibility of L2 production. A slightly larger effect size for experienced raters ($g = .44$) compared to inexperienced raters ($g = 0.36$) may point to a higher consistency for experienced raters (Isaacs & Thomson, 2013). However, the small and nonsignificant difference suggests that the choice of experienced or inexperienced raters does not appear to substantially affect the evaluation of L2 production accuracy. Both experienced and less experienced raters were equally reliable. The number of talkers during training was not a significant predictor, suggesting that increasing or decreasing talker variability within the range of 3 to 30 may not affect the transfer of HVPT to production. A minimum of 3 talkers may be sufficient for significant production gains to be observed. However, the majority of studies limited the number of talkers from 3 to 6 ($k = 37$); thus, talker effects with greater numbers need to be explored in future research before the optimal number of talkers is determined. On the other hand, other variables related to learner profiles (learning context and AOT), training features (phonetic information, duration, and training time per session) were found to have a major impact on production gains. Further, choice of assessment for production tests (outcome measures, elicitation tasks, and prompt modality) resulted in differential findings. In what follows, findings of these key variables (learning context, AOT, phonetic information, training duration, training time per session, and assessment) will be discussed in greater detail.

Learner features

Regarding the context of learning, the relatively larger effect for FL ($g = 0.60$) compared to SL ($g = 0.34$) indicates that HVPT was likely to be less impactful on L2 production for SL learners than FL learners. The smaller benefit of perception training for SL learners may be attributed to their higher degree of L2 experience developed through immersion in naturalistic settings (Iverson et al., 2012). Eleven studies reporting information about the mean length of residence showed that SL learners in our meta-analysis were considered relatively experienced in terms of the length of residence in the target language country ($M = 18$ months, range $= 1–49$ months). According to the hypothesis of a window of maximal opportunity (Derwing & Munro, 2015), substantial phonetic learning occurs during the early period of residence in the L2-speaking country. Thus, it is possible that SL learners' perception begins to plateau after a period of immersion, or at least further increases are so modest (e.g., Derwing & Munro, 2013) that it would take much longer to detect change than the short durations afforded by the studies considered here. Perceptual categories that have already changed and stabilized may not improve as much from perception training when compared with less developed representations of FL learners, which may cause perception training to be less impactful on SL learners' production (Nagle & Baese-Berk, 2022).

The negative relationship between AOT and production gains indicated that older adult learners tended to benefit less from HVPT than younger adult learners in the current data of participants whose age ranged from 12 to 37. One possible reason for this finding is that it may be more challenging for learners with established L1 perceptual knowledge to improve L2 production accuracy. Given that older learners are assumed to receive a greater amount of L1 input and develop a

more robust L1 categorical knowledge (Flege, 1995a), the interference from L1 knowledge with L2 category learning can be considered larger. As the L1 becomes more robust, older learners may have more difficulty discriminating phonetic differences that exist crosslinguistically, which might reduce the beneficial impact of training on production outcomes. It is important to note, however, that the mean effect of age does not mean that older learners cannot benefit. For example, the oldest learner (fifty years old) in Thomson (2011) evidenced among the greatest improvement in production of the participants in that study.

Perception training features

Provision of phonetic information had a positive impact on the training effect on production gains ($g = 0.94$) compared to no phonetic information ($g = 0.45$). These findings support the meta-analysis of general perception training on production improvement (Sakai & Moorman, 2018) showing that a large effect was observed for learners presented with phonetic information ($d = 0.99$) compared to those without such information ($d = 0.40$). Explicit attention to and knowledge of target forms are likely to contribute to the development of L2 production accuracy (Saito, 2019), and drawing learners' attention to target forms increases the efficacy of L2 phonetic training (Guion & Pederson, 2007). The robustness of these findings suggests an important pedagogical implication for L2 speech learning as presenting phonetic information about target sounds prior to having learners complete perception training can optimize the transfer of HVPT to production. This may be due to their ability to apply explicit knowledge to production, particularly in highly controlled speaking tasks, which predominated the studies included in this meta-analysis.

Regarding training duration, training that lasted for one month or longer ($g = 0.32$) was less effective than training that lasted for shorter than one month ($g = 0.61$). Scheduling training programs for a longer period of time may not necessarily bring about the most benefit for production learning (Wong, 2013). A recent meta-analysis by Kim and Webb (2022) suggests that a shorter spacing interval during learning is particularly beneficial, in light of the high degree of complexity involved in the learning of pronunciation. Given that studies scheduling a longer-term training program tend to provide longer intervals between training (e.g., 8 sessions within 2 weeks in Lee & Lyster, 2017 vs. 5 sessions within 5 weeks in Carlet, 2017), in such studies, learners might have been exposed to training stimuli with a longer spacing interval between training sessions. When the spacing is longer, learners may have difficulty accessing phonological information during subsequent exposures to auditory input, which might have reduced the efficacy of perception training.

Training time per session was negatively associated with production learning, indicating that the longer the time per training session was, the lower the production gain was. Despite the significant association, the scatterplot for this relationship (see Figure 6) showed variability with several studies deviating from the regression line. This figure appears to reveal that studies around 25 minutes consistently produce relatively higher effect sizes ($g = 0.44$ to 1.62) after which effect sizes seemed to drop at around 30 minutes ($g = 0.17$ to 0.96) and afterward with a few exceptions. The data tentatively suggest that perhaps 20 to 30 minutes may be appropriate for
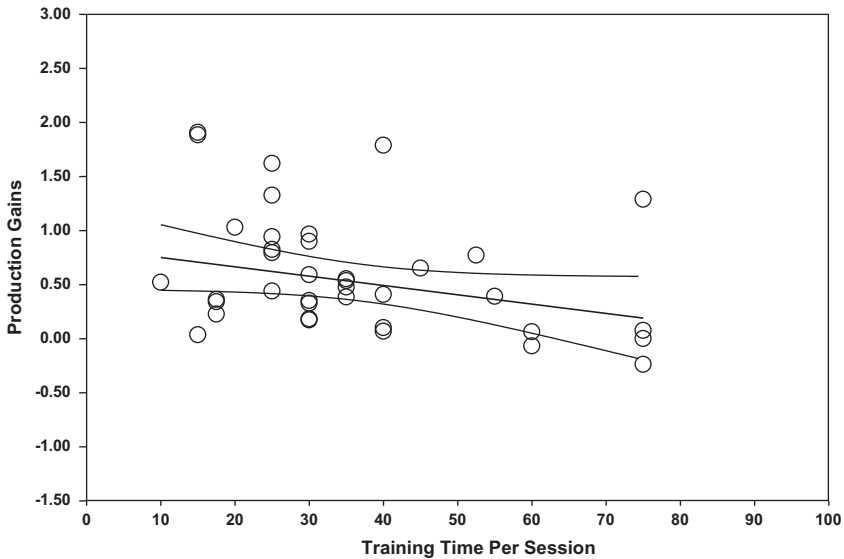
Figure 6. Scatterplot of training time per session (min) and production gains (Hedges' *g*).

the amount of training time per session in order to maximize the training benefit for production learning. Given that phonetic training tends to be tedious without any incentive (e.g., giving monetary rewards for correct responses in Bradlow et al., 1999), training for 30 minutes or longer may not necessarily increase the effect of training (Logan & Pruitt, 1995).

Features of production tests

The moderator analysis of outcome measures showed a larger training effect, in the order of transcription ($g = 1.14$), identification ($g = 0.45$), human rating ($g = 0.28$), and acoustic analysis ($g = 0.18$). The nonsignificant and small effect for acoustic analysis is not consistent with previous meta-analyses of L2 speech learning (Sakai & Moorman, 2018; Saito & Plonsky, 2019), but the result needs to be interpreted with caution due to small sample size ($k = 3$). A relatively large effect size for transcription from small sample size ($k = 4$) also needs to be interpreted with caution. The result may be confounded by other moderator variables. Particularly, the large effect was mainly attributed to two effect sizes ($g > 2.0$) from a single study (Wong, 2013), which focused on younger learners ($M = 16.3$ and 16.4 years old vs. overall mean across all studies $= 23.9$ years old) in the FL context and the phonetic training lasted for 15 minutes, shorter than the overall mean across all studies ($M = 35.8$ minutes).

A smaller effect was observed for human ratings compared to intelligibility measured by native speaker forced-choice identification. To further examine the small effect for human ratings, we examined the rating scales used in 11 relevant studies adopting scalar rating measures. As summarized in Table 2, researchers measured three constructs of L2 pronunciation proficiency in terms of general

Table 2. Summary of 11 studies using human rating to measure production accuracy of L2 sounds

Study	Construct	Scale points	Scale labels	g
Carlet (2017)	Comprehensibility	9	1 = hard to identify as the selected sound, 9 = easy to identify as the selected sound	0.49
Dong et al. (2019)	Nativelikeness	7	1 = not recognizable, 7 = native speaker level	−0.09
Hardison (2003)	Accuracy	7	7 = a good example of the target	0.90
Hazan et al. (2005)	Accuracy	7	1 = bad, 7 = excellent	0.18
Hwang & Lee (2015)	Nativelikeness	7	1 = not at all native-like, 7 = native-like	0.10
Lee & Lyster (2017)	Comprehensibility	5	1 = difficult to understand, 5 = easy to understand	0.28
Lopez-Soto & Kewley-Port (2009)	Nativelikeness & Comprehensibility	3	1 = extremely poor pronunciation, 2 = strong foreign accent, 3 = acceptable pronunciation (even if a foreign accent is perceived)	0.07
MacDonald (2012)	Nativelikeness	7	1 = very accurate/native-like, 7 = very inaccurate/clearly not native	0.13
Reyes & Hazan (2021)	Accuracy	7	1= very poor, 7 = very good	0.77
Thomson & Derwing (2016)	Accuracy	3	0 = another category, 1 = poor, 2 = good	0.13
Trakantalerngsak (2016)	Nativelikeness	5	1 = a poor exemplar of the target consonant or heavily accented or not Japanese native-like, 5 = a good, Japanese-sounding attempt or native-like or not accented at all	0.34

Note: Multiple effect sizes were averaged to yield a single effect size per study. Some researchers stated they measured "intelligibility" but it seems to conflate it with nativelikeness (e.g., Dong et al., 2019; Trakantalerngsak, 2016).

accuracy (e.g., correct vs. incorrect), nativelikeness (i.e., how different L2 speech is from an L1 variety), and comprehensibility (i.e., how easily L2 speech is identified). When general accuracy was further specified in the manuscript or rating labels (e.g., Macdonald, 2012 for nativelikeness), such studies were assigned to the specific construct (nativelikeness or comprehensibility). In one study (Lopez-Soto & Kewley-Port, 2009), the scale descriptors focused on both nativelikeness and comprehensibility in the same scale, which was thus labeled as nativelikeness & comprehensibility. The analysis of average effect size descriptively demonstrates that the training effect tends to be smaller when nativelikeness rating was used ($g = 0.09$) compared to comprehensibility ($g = 0.36$) or accuracy ($g = 0.39$). The relatively smaller effect of training for nativelikeness compared to comprehensibility

and intelligibility is consistent with earlier studies showing that L2 oral production becomes more comprehensible and intelligible with continued exposure to L2 spoken input (Derwing & Munro, 2013; Uchihara et al., 2022) and explicit instruction (Derwing et al., 1998), whereas development of native-like pronunciation tends to be slow and gradual (Saito, 2015). Accordingly, perhaps the relatively smaller effect for human rating in this meta-analysis may have been attributed to the inclusion of studies focusing on nativelikeness.

Elicitation tasks and prompt modality

Regarding the production elicitation tasks, inconsistency in the effect of training depending on the assessment task used to measure production indicates that the tasks are incommensurate and likely measuring different phenomena. A much larger effect of training was observed in the order of word reading ($g = 0.72$), recalling ($g = 0.58$), sentence reading ($g = 0.45$), and repetition ($g = 0.28$). These findings can be explained in terms of varied cognitive demands present across tasks. Word reading is the least cognitively demanding and allows learners to direct their explicit attention to target sounds; hence, production gains as evaluated by a word reading task might be tapping into changes in learners' declarative knowledge rather than procedural knowledge (Thomson, 2022a; Thomson & Derwing, 2015). In contrast, a word recall task may involve a slightly higher degree of retrieval effort and pronunciation may be at least partially influenced by procedural knowledge. Similarly, sentence reading tasks also provide greater distraction than word reading tasks, allowing less reference to declarative knowledge. In contrast, a delayed repetition task forces learners to move phonetic information obtained from the prompt into long-term memory and then retrieve it after an interruption. This means that the resulting production is more likely to be an indication of their developing interlanguage system. The effect of differential demands on attention on production accuracy was confirmed by the result that prompts providing only auditory input led to a smaller effect ($g = 0.19$) than when orthographic input is provided ($g = 0.62$ for written input; $g = 0.56$ for auditory + written input). Recalling and repetition tasks are similar in many respects except for the provision of auditory input. In terms of cognitive load, a similar degree of cognitive demands may be required from both recalling (e.g., presented with a card on which a test word is printed, then turned over, and asked to recall and pronounce the word in Hardison, 2003) and delayed repetition (e.g., pronounce after white noise in Dong et al., 2019 or after 3000ms in Aliaga-García, 2017). Despite this similarity, a much smaller effect was found for repetition ($g = 0.28$) compared to recalling ($g = 0.58$), suggesting that the exposure to auditory stimuli, regardless of efforts to mitigate the influence of spoken input, may induce a high degree of phonological involvement with target sounds (Llompart & Reinisch, 2019) and impact production outcomes.

Ultimately, we take the view that the choice of assessment should be determined by the goal of the assessment. If the goal is to test declarative knowledge, then reading tasks and other highly controlled tasks will work well. If, however, the goal is to evaluate the extent to which changes in perception manifest in changes in procedural knowledge in production, then less controlled tasks, such as delayed repetition or even spontaneous speaking tasks should be used. The latter should be

considered the gold standard, while the former's usefulness may be primarily limited to formative assessment. This task hierarchy is evident in Thomson and Derwing's (2016) HVPT perceptual training study in which they found improvement in production in a delayed imitation task, but not in a more spontaneous production task which elicited the same targets in sentences that the learners were asked to create on the fly.

Conclusion

The primary goal of the current meta-analysis was to examine to what extent HVPT brings about perception-to-production transfer in terms of gains, retention, and generalization, and which variables related to learner, training, and outcome measures modulate the overall effectiveness of HVPT. The present meta-analysis provided partial support for a relationship between L2 speech perception and production. It provides evidence for the transfer of HVPT to production with small-to-medium effects and further supports the significant relationship between perception and production learning and retention at the study level. However, we did not find strong support for long-term retention of production learning and generalization to untrained stimuli, nor a significant perception-production link for gain scores at the participant level. In addition, several learner-related variables were found to influence production gains, and methodological choices regarding how to assess production also impacted the magnitude and/or type of gain evidenced. These findings provided insights into various factors contributing to the efficacy of HVPT for L2 speech production learning and its measurement.

With the aim of optimizing the effectiveness of HVPT in improving L2 production, the findings of this meta-analysis have several pedagogical implications. First, the weak evidence for the production transfer in this meta-analysis suggests that delaying the introduction of explicit production training may be worth considering. It is possible that without more finely tuned perceptual categories, attempts to match production against emergent perceptual categories break down. The learners' imprecise productions would become perceptual input, leading to the erosion of perceptual gains (Thomson, 2022b). As such, there could be a benefit of delaying production practice until the perceptual system is capable of self-monitoring production. However, it should be noted that the current meta-analysis exclusively focused on perception training without production training; thus, the extent to which an additional component of production practice promotes (or hinders) the transfer remains unknown. We hope that, as the studies increase in number, a future meta-analysis will focus on production HVPT studies while considering a variety of training features revealed by some successful cases, such as combinational effects (perception+ production vs. production; Herd et al., 2013), training task (e.g., imitation task; Aliaga-García, 2017), visual input (waveform inspection; Herd et al., 2013), and stimuli status (nonwords vs. real words; Mora et al., 2022).

The current findings also offer practical guidance with the goal of optimizing the production transfer. Practitioners should note that production learning is enhanced

when HVPT involves (a) a focused and intensive training program (e.g., 20 to 30 min sessions for less than one month), (b) provision of explicit phonetic information that promotes production learning, (c) younger L2 learners, and (d) learners studying L2 in foreign language contexts. It is also important to note that HVPT is better suited to improve explicit knowledge related to L2 speech production (elicited via word reading), while whether it improves procedural knowledge needed for spontaneous speech production remains uncertain.

Lastly, there are several suggestions for future HVPT studies with a view to improving methodological soundness and rigor in the investigation of the perception-production link. First, based on the finding that the average sample size was low ($N = 17.06$), future HVPT research should endeavor to have more participants (for the same call, see also Sakai & Moorman, 2018). Given the great amount of time and cost required for the implementation of HVPT experiments, researchers may find online experiment builders such as Gorilla (Anwyl-Irvine et al., 2020) or online training platforms (e.g., English Accent Coach, Thomson, 2012a) useful. Second, based on the findings that interrater reliability was not reported in many studies (unreported rates: 57% for native speaker identification, 38% for scalar ratings, 50% for transcription), we urge that future studies should report the pretest and posttest interrater agreement and evaluate the impact of test reliability on their results. Third, spontaneous tasks should be used to elicit L2 speech production. In this meta-analysis, all studies predominantly relied on using controlled production tasks (reading aloud, recalling, and repetition), reflecting the same trend in the L2 pronunciation literature in general (Thomson & Derwing, 2015). Using spontaneous tasks such as picture naming (e.g., Nagle, 2021) is important from a perspective of ecological validity and will provide additional insights into the multidimensional view of the perception-production link. Fourth, the construct definition of production accuracy needs to be specified in adopting scalar rating measurements. In some studies, both features of nativelikeness and comprehensibility or intelligibility were mentioned in a single descriptor or rating label. Given that the constructs of nativelikeness, comprehensibility, and intelligibility are confirmed to be independent at the sentence (Derwing & Munro, 2009) or word (Uchihara, 2022) level, it is advisable to treat them separately in measuring global constructs of L2 pronunciation proficiency. Further, general accuracy (i.e., correct vs. incorrect distinctions) may need to be defined more clearly. Such a broad description of L2 pronunciation accuracy may cause variations in the way listeners interpret what constitutes accuracy, given that the intended message may not be clear—"was the intention 'close enough to be recognized as a particular phoneme' or was it 'native-like versus non-native-like'?" (Thomson & Derwing, 2015, p. 337).

Notes

1 No significant difference in effect size was found between test-only and nontarget-training conditions, $Q(1) = 0.74$, $p = .389$.

2 No significant difference was found between audiovisual and audio-only training conditions, $Q(1) = 2.57$, $p = .109$.

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 [Unpublished doctoral dissertation]. Universitat de Barcelona.

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

Baker, W., & Trofimovich, P. (2006). Perceptual paths to accurate production of L2 vowels: The role of individual differences. *International Review of Applied Linguistics in Language Teaching*, **44**(3), 231–250. https://doi.org/10.1515/IRAL.2006.010

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

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In M. J. Munro & O.-S. Bohn (Eds.), *Second language speech learning: The role of language experience in speech perception and production* (pp. 13–24). John Benjamins.

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

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

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

Carlet, A., & Cebrian, J. (2022). The roles of task, segment type, and attention in L2 perceptual training. *Applied Psycholinguistics*, **43**(2), 271–299. https://doi.org/10.1017/S0142716421000515

Casillas, J. V. (2020). Phonetic category formation is perceptually driven during the early stages of adult L2 development. *Language and Speech*, **63**(3), 550–581. https://doi.org/10.1177/0023830919866225

Counselman, D. (2010). Improving pronunciation instruction in the second language classroom (Unpublished doctoral dissertation).

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, **42**(4), 476–490. https://doi.org/10.1017/S026144480800551X

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning*, **63**(2), 163–185. https://doi.org/10.1111/lang.12000

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

Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning*, **48**(3), 393–410. https://doi.org/10.1111/0023-8333.00047

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, **54**(4), 655–679. https://doi.org/10.1111/j.1467-9922.2004.00282.x

Dong, H., Clayards, M., Brown, H., & Wonnacott, E. (2019). The effects of high versus low talker variability and individual aptitude on phonetic training of Mandarin lexical tones. *PeerJ*, **7**, e7191. https://doi.org/10.7717/peerj.7191

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

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

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

Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, **25**(4), 437–470. https://doi.org/10.1006/jpho.1997.0052

Flege, J. E., MacKay, I. R. A., & Meador, D. (1999). Native Italian speakers' perception and production of English vowels. *The Journal of the Acoustical Society of America*, **106**(5), 2973–2987. https://doi.org/10.1121/1.428116

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

Ghaffarvand Mokari, P., & Werner, S. (2018). Perceptual training of second-language vowels: Does musical ability play a role? *Journal of Psycholinguistic Research*, **47**(1), 95–112. https://doi.org/10.1007/s10936-017-9517-8

Guion, S., & Pederson, E. (2007). Investigating the role of attention in phonetic learning. In O.-S. Bohn & M. Munro (Eds.), *Second-language speech learning: The role of language experience in speech perception and production: A festschrift in honour of James E. Flege* (pp. 57–77). John Benjamins.

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

Hattori, K., & Iverson, P. (2009). English /r/-/l/ category assimilation by Japanese adults: Individual differences and the link to identification accuracy. *The Journal of the Acoustical Society of America*, **125**(1), 469–479. https://doi.org/10.1121/1.3021295

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

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

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

Hwang, H., & Lee, H.-Y. (2015). The effect of high variability phonetic training on the production of English vowels and consonants. In the Scottish Consortium for ICPhS (Ed.), Proceedings of the 18th international Congress of Phonetic Sciences (Paper number 486). Glasgow, UK: Glasgow University.

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, **10**(2), 135–159. https://doi.org/10.1080/15434303.2013.769545

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

Jia, G., Strange, W., Wu, Y., Collado, J., & Guan, Q. (2006). Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2 exposure. *The Journal of the Acoustical Society of America*, **119**(2), 1118–1130. https://doi.org/10.1121/1.2151806

Kartushina, N., & Frauenfelder, U. H. (2014). On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation. Frontiers in Psychology, 5, 105122. https://www.frontiersin.org/articles/10.3389/fpsyg.2014.01246

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2015). The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. *The Journal of the Acoustical Society of America*, **138**(2), 817–832. https://doi.org/10.1121/1.4926561

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *The Canadian Modern Language Review*, **64**(3), 459–489. https://doi.org/10.3138/cmlr.64.3.459

Kim, S. K., & Webb, S. (2022). The effects of spaced practice on second language learning: A meta-analysis. *Language Learning*, **72**(1), 269–319. https://doi.org/10.1111/lang.12479

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

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

Lee, A. H., & Lyster, R. (2017). Can corrective feedback on second language speech perception errors affect production accuracy? *Applied Psycholinguistics*, **38**(2), 371–393. https://doi.org/10.1017/S0142716416000254

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, **36**(3), 345–366. https://doi.org/10.1093/applin/amu040

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

Li, Y. (2015). Audio-visual training effect on L2 perception and production of English /θ/-/s/ and /ð/-/z/ by Mandarin speakers [Unpublished doctoral dissertation]. Newcastle University.

Llompart, M., & Reinisch, E. (2019). Imitation in a second language relies on phonological categories but does not reflect the productive usage of difficult sound contrasts. *Language and Speech*, **62**(3), 594–622. https://doi.org/10.1177/0023830918803978

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

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

Lopez-Soto, T., & Kewley-Port, D. (2009). Relation of perception training to production of codas in English as a Second Language. *Proceedings of Meetings on Acoustics*, 6(1), 062003. https://doi.org/10.1121/1.3262006

Macdonald, R. M. M. (2012). Counteracting age related effects in L2 acquisition: Training to distinguish between French vowels [Unpublished doctoral dissertation]. The University of Edinburgh.

Melnik-Leroy, G. A., Turnbull, R., & Peperkamp, S. (2022). On the relationship between perception and production of L2 sounds: Evidence from Anglophones' processing of the French /u/–/y/ contrast. *Second Language Research*, **38**(3), 581–605. https://doi.org/10.1177/0267658320988061

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

Nagle, C. L. (2018). Examining the temporal structure of the perception–production link in second language acquisition: A longitudinal study. *Language Learning*, **68**(1), 234–270. https://doi.org/10.1111/lang.12275

Nagle, C. L. (2021). Revisiting perception–production relationships: Exploring a new approach to investigate perception as a time-varying predictor. *Language Learning*, **71**(1), 243–279. https://doi.org/10.1111/lang.12431

Nagle, C. L., & Baese-Berk, M. M. (2022). Advancing the state of the art in L2 speech perception-production research: Revisiting theoretical assumptions and methodological practices. *Studies in Second Language Acquisition*, **44**(2), 580–605. https://doi.org/10.1017/S0272263121000371

Nearey, T. M. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of America*, **101**(6), 3241–3254.

Okuno, T., & Hardison, D. M. (2016). Perception-production link in L2 Japanese vowel duration: Training with technology. *Language Learning & Technology*, **20**(2), 61–80.

Peperkamp, S., & Bouchon, C. (2011). The relation between perception and production in L2 phonological processing. *Proceedings of Interspeech*, **12**, 161–164.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency effects and the emergence of linguistic structure* (pp. 137–157). John Benhamins.

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

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

Reyes, Y. P., & Hazan, V. (2021). English vowel perception by non-native speakers: Impact of audio and visual training modalities. *Onomázein*, **51**, 112–136. https://doi.org/10.7764/onomazein.51.04

Saito, K. (2015). Experience effects on the development of late second language learners' oral proficiency. *Language Learning*, **65**(3), 563–595. https://doi.org/10.1111/lang.12120

Saito, K. (2019). Individual differences in second language speech learning in classroom settings: Roles of awareness in the longitudinal development of Japanese learners' English /ɹ/ pronunciation. *Second Language Research*, **35**(2), 149–172. https://doi.org/10.1177/0267658318768342

Saito, K. (2021). What characterizes comprehensible and native-like pronunciation among English-as-a-second-language speakers? Meta-analyses of phonological, rater, and instructional factors. *TESOL Quarterly*, **55**(3), 866–900. https://doi.org/10.1002/tesq.3027

Saito, K., Kachlicka, M., Suzukida, Y., Petrova, K., Lee, B. J., & Tierney, A. (2022). Auditory precision hypothesis-L2: Dimension-specific relationships between auditory processing and second language segmental learning. *Cognition*, **229**, 105236. https://doi.org/10.1016/j.cognition.2022.105236

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

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, **38**(4), 439–462. https://doi.org/10.1093/applin/amv047

Saito, K., & van Poeteren, K. (2018). The perception–production link revisited: The case of Japanese learners' English /ɹ/ performance. *International Journal of Applied Linguistics*, **28**(1), 3–17. https://doi.org/10.1111/ijal.12175

Sakai, M., & Moorman, C. (2018). Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research. *Applied Psycholinguistics*, **39**(1), 187–224. https://doi.org/10.1017/S0142716417000418

Scovel, T. (1969). Foreign accents, language acquisition, and cerebral dominance. *Language Learning*, **19**(3-4), 245–253. https://doi.org/10.1111/j.1467-1770.1969.tb00466.x

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

Silpachai, A. (2020). The role of talker variability in the perceptual learning of Mandarin tones by American English listeners. *Journal of Second Language Pronunciation*, **6**(2), 209–235. https://doi.org/10.1075/jslp.19010.sil

Soler-Urzúa, F. (2011). The acquisition of English /l/ by Spanish speakers via text-to-speech synthesizers: A quasi-experimental study (Unpublished master's thesis).

Thomson, R. I. (2008). L2 English vowel learning by Mandarin speakers: Does perception precede production? *Canadian Acoustics*, **36**(3), 134–135. Proceedings of the annual conference of the Canadian Acoustics Association.

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

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

Thomson, R. I. (2012b). Improving L2 listeners' perception of English vowels: A computer-mediated approach. *Language Learning*, **62**(4), 1231–1258. https://doi.org/10.1111/j.1467-9922.2012.00724.x

Thomson, R. I. (2018). High Variability [Pronunciation] Training (HVPT): A proven technique about which every language teacher and learner ought to know. *Journal of Second Language Pronunciation*, **4**(2), 208–231. https://doi.org/10.1075/jslp.17038.tho

Thomson, R. I. (2022a). The relationship between L2 speech perception and production. In T. M. Derwing, M. J. Munro & R. I. Thomson (Eds.), *The Routledge handbook of second language acquisition and speaking* (pp. 372–385). Routledge.

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

Thomson, R. I., & Derwing, T. M. (2015). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, **36**(3), 326–344. https://doi.org/10.1093/applin/amu076

Thomson, R. I., & Derwing, T. M. (2016). Is phonemic training using nonsense or real words more effective? In J. Levis, H. Le, I. Lucic, E. Simpson & S. Vo (Eds.), Proceedings of the 7th pronunciation in second language learning and teaching conference (pp. 88–97). Iowa State University.

Thomson, R. I., & Isaacs, T. (2009). Within-category variation in L2 English vowel learning. *Canadian Acoustics*, **37**(3), 138–139.

Trakantalerngsak, T. (2016). The effect of perceptual training on the learning of Japanese fricatives and affricate contrasts by native Thai learners of Japanese [Unpublished doctoral dissertation]. Osaka University.

Uchihara, T. (2022). Is it possible to measure word-level comprehensibility and accentedness as independent constructs of pronunciation knowledge? *Research Methods in Applied Linguistics*, **1**(2), 100011. https://doi.org/10.1016/j.rmal.2022.100011

Uchihara, T., Karas, M., & Thomson, R. I. (2021). *High Variability Phonetic Training (HVPT): A meta-analysis* [Paper presentation]. St. Catharines, ON, Canada: Paper presented at Pronunciation in Second Language Learning and Teaching (PSLLT) Conference.

Uchihara, T., Webb, S., Saito, K., & Trofimovich, P. (2022). Frequency of exposure influences accentedness and comprehensibility in learners' pronunciation of second language words. *Language Learning.* https://doi.org/10.1111/lang.12517

Underbakke, M. E. (1993). Hearing the different: Improving Japanese students' pronunciation of a second language through listening. *Language Quarterly*, **31**, 67–89.

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America*, **113**(2), 1033–1043. https://doi.org/10.1121/1.1531176

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America*, **106**(6), 3649–3658. https://doi.org/10.1121/1.428217

Wong, W. S. (2013). Training the perception and production of English vowels /I/-/iː/, /e/-/æ/ and /ʊ/-/uː/ by Cantonese ESL learners in Hong Kong [Unpublished doctoral dissertation]. The Chinese University of Hong Kong.

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

Zhang, X., Cheng, B., & Zhang, Y. (2021). The role of talker variability in nonnative phonetic learning: A systematic review and meta-Analysis. *Journal of Speech, Language, and Hearing Research*, **64**(12), 4802–4825. https://doi.org/10.1044/2021_JSLHR-21-00181