Applied Psycholinguistics **39** (2018), 187–224
doi:10.1017/S0142716417000418

Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research

MARI SAKAI and COLLEEN MOORMAN
Georgetown University

Received: July 5, 2016   Accepted for publication: August 11, 2017

ADDRESS FOR CORRESPONDENCE
Mari Sakai, Graduate Programs, Georgetown University Law Center, 600 New Jersey
Avenue NW, Washington, DC 20001. E-mail: ms2335@georgetown.edu

ABSTRACT
Cognitive scientists across disciplines have shown a vested interest in examining if and how the speech perception and production modalities are connected. The field of second language (L2) acquisition contributes to this discussion by investigating the effects of auditory training of L2 sounds on pronunciation. This meta-analysis offers a comprehensive view of the last 25 years of L2 perception training studies that test for effects in production. The results indicate that the two modalities are connected, insomuch as training the perception of L2 sounds can induce positive change in the productive mode as well. The data indicate that strictly controlled perception training led to medium-sized improvements in perception ($d = 0.92$, $SD = 0.96$) and small improvements in production ($d = 0.54$, $SD = 0.45$). A correlation analysis suggests a small- to medium-sized relationship between perception and production gains, although this relationship was not significant. The production of obstruents improved to a larger degree than sonorants or vowels, and an additional six moderating variables influenced the magnitude of the production effect sizes. We caution researchers to not equate the connection of the two modalities in long-term linguistic development to real-time neurological processing, and we end with five recommendations for the domain of L2 phonetic training research.

Keywords: meta-analysis; perception training; second language; speech learning model

In 1988, Jim Flege began a series of studies that investigated the acquisition of second language (L2) speech sounds (Flege, 1988a, 1988b, 1992, 1995a, 1995b, 2002, 2003), which formed the foundation of what is now known as the speech learning model (SLM). Among other postulates and hypotheses, this theory of speech learning states that (a) perception of L2 sounds can improve after sufficient

© Cambridge University Press 2017 0142-7164/17

exposure and (b) accurate perception is a necessary precursor to targetlike production (Flege, 1995a). After the introduction of this model, the number of phonological perception training studies in the field of second language acquisition (SLA) increased dramatically, as researchers began to test the SLM postulates (e.g., Herd, Jongman, & Sereno, 2013; Iverson & Evans, 2009). The results were consistent that adult participants were able to improve their perception of L2 sounds after sufficient training.

The second hypothesis described here places a large emphasis on the interconnected nature of the perception and production modalities. The SLM states that there is an intricate and detailed auditory-to-articulatory mapping that connects what language users hear to what they produce, and that nonnative speakers' productions will eventually reflect the mental representations of the L2 phonemes (Flege, 1995a, 2003). If the mental representation is created by the perception of distinctive features of the L2 phoneme, and production mirrors the mental representation, then production is constrained by perceptual abilities. Flege (2003) asserted, "L2 phonetic segments cannot be produced accurately unless they are perceived accurately" (p. 27). This claim implies that a nonnative speaker's production accuracy will not precede or lead perceptual accuracy. Furthermore, it is not expected that productive abilities will be more nativelike than segmental perception (p. 26). In other words, the SLM predicts a unidirectional relationship for the acquisition of new sounds, from the perceptual mode through the mental representation to the production mode.

Researchers in an array of language-related fields have shown great interest in understanding the nature and complexities of the connection between the perception and production modalities. Experiments in first language (L1) acquisition and the deaf and hard of hearing population have investigated development patterns and interactions between the two modalities at one moment in time or on a longitudinal scale. The literature in L1 perception and production research is expansive (e.g., Byun, 2015; Rvachew, Nowak, & Cloutier, 2004; Wong, Schwartz, & Jenkins, 2005; see Geravain & Mehler, 2010; Meyer, Huettig, & Levelt, 2016; and Werker & Hensch, 2015, for reviews of perception and production in L1 development). One example of literature in this field is Altvater-Mackensen and Fikkert (2010), who documented the productive development of obstruents in Dutch-speaking toddlers and also tested the children's perceptual ability of the same sounds. They found an asymmetrical development for the production and perception of obstruents in different parts of a word (i.e., onsets and codas). Another L1 example is Byun (2012), who conducted a case study of 4-year-olds with signs of phonological delay. She asked which modality caused neutralization problems in the child's speech: did the production errors lead to perception errors or vice versa? The results evidenced that the production neutralization fed perceptual problems. In the deaf and hard of hearing population, researchers are concerned with the question of productive development in an imperfect auditory environment. Many investigators have run correlations between the amount of hearing loss and the intelligibility of speech, and also perception and production development after children receive a cochlear implant (e.g., Golfeto & de Souza, 2015; Song et al., 2015; Tseng, Kuei, & Tsou, 2011).

https://doi.org/10.1017/S0142716417000418 Published online by Cambridge University Press

In addition to the L1 and deaf and hard of hearing literature that present the relationship of the modalities in a long-term context, experiments in cognitive neuroscience offer an abundance of information about online or simultaneous processing of the modalities in a window of seconds or milliseconds. The classic view in neurocognition posits auditory and productive speech processing occur in largely different systems and different areas of the brain: hearing speech sounds activates the superior temporal gyrus in the left hemisphere, while producing speech sounds activates the motor cortex in the right hemisphere. However, over the last decade, many neuroimaging studies have begun to find overlap, as passive speech listening activates the motor cortex. A study conducted by Wilson, Saygin, Sereno, and Iacoboni (2004) scanned 10 participants in a functional magnetic resonance imaging study as they listened to and later repeated monosyllabic nonwords. During the perception portion of the task, all 10 participants showed activation in the ventral premotor cortex spanning to the primary motor cortex. Brain regions that were activated during the production task overlapped with the perception task by 73%. The researchers concluded, "These findings are consistent with the view that speech perception involves the motor system in a process of auditory-to-articulatory mapping across a phonetic code with motor properties" (p. 702). In the years following this study, others have corroborated the findings that the motor cortex is involved in passive listening of speech and singing (e.g., Chen, Penhune, & Zatorre, 2008; D'Ausilio, Bufalari, Salmas, & Fadiga, 2012; Skipper, Nusbaum, & Small, 2005). Neuroimaging research has been able to show the opposite direction as well. Similar to the methodological design in Wilson et al. (2004), Calvert et al. (1997) and Campbell et al. (2001) showed that participants' auditory and motor cortices were activated when they viewed silent videos of people mouthing speech sounds. This bilateral activation was not found when viewing nonspeech mouth movements. In sum, functional brain imaging provides evidence of a bidirectional influence between the perception and production modalities in real-time processing. Again, neuroscience investigates the immediate and simultaneous actions involved in speech perception and production, while L1 and the deaf and hard of hearing studies add information about the long-term interaction of the modalities.

The field of SLA has been able to contribute to this discussion in at least two ways, by (a) running correlation analyses on adults' L2 perception and production abilities and (b) training adult learners in the perception mode and testing for gains in the production mode. The latter type of experimental design is valuable because it can offer strong evidence of the two modalities being connected if training in the perception mode results in improvements in the production mode. Because these studies have occurred across many language combinations and with a large variety of target phonemes, it has been difficult for researchers to compile the results of all of these studies, and currently the picture is quite fragmented. In order to address the present state of knowledge about the perception–production link, this meta-analysis aims to bring together results from all L2 perceptual training studies that have tested for production gains. The results have the potential to add valuable information to the cross-disciplinary understanding of the complex nature of the perception–production connection.

THE DIFFICULTY OF LEARNING NEW SOUNDS

One of the challenges that adults face as they learn an L2 is the acquisition of new phonemes, in terms of both the creation of accurate mental representations of these phonemes and the ability to perceive and produce them. Proponents of the critical period hypothesis argue that older learners will not be able to attain complete nativelike phonetic knowledge (e.g., Granena & Long, 2013; Lenneberg, 1967; Long, 1990; Patkowski, 1994; Stölten, Abrahamsson, & Hyltenstam, 2015). By contrast, Flege's (1995a) SLM states that adults do retain the ability to create new phonetic categories for L2 sounds. However, in the process of acquiring nonnative phonetic systems, learners undergo an interlanguage period where their first language (L1) categories mitigate the acquisition of the L2 phonemes. As Cenoz and García Lecumberri (1999) explain, adult learners "rely on phonetic rather than sensory perception so that their perception of L2 sounds is biased by their L1 phonetic system and [learners] tend to perceive L2 sounds in terms of the categories in the L1" (p. 262). In cases where a phone from the L2 is discernable as a distinct sound from the L1 phonemes, a new mental representation will successfully be added to the native phonemic system. However, in instances when a learner cannot discriminate differential features of an L2 phone, the sound assimilates to the closest existing L1 representation to create a composite category, which carries information about both the L1 and the L2 phonemes (Flege, 1995a). In the presence of enough naturalistic exposure, distinct L2 phonetic representations can be created on their own. However, with a lack of sufficient input, this phenomenon of categorical assimilation may be one cause of L2 perception errors, and may be the reason some phones persist in causing confusion and difficulty.

PERCEPTUAL TRAINING OF PHONEMES AND THE CONNECTION TO SPEECH PRODUCTION

The difficulties that nonnative speakers encounter with perception of particular phonemic contrasts present an appealing environment for researchers interested in the plasticity of an adult learner's native phonemic representations (e.g., de Leeuw, Mennen, & Scobbie, 2013; Hazan, Sennema, Iba, & Faulkner, 2005; Iverson, Pinet, & Evans, 2012; Lambacher, Martens, & Kakehi, 2002; Lengeris, 2008; Lively, Pisoni, & Logan, 1991; Mora & Nadeu, 2012). In cases where naturalistic exposure is not sufficient to help the learner create a new phoneme category, training can provide enough dense and targeted input to help the learner perceive the problematic phone(s) correctly.

Motivated by this hypothesis, many researchers have explored the effects of distinct perceptual trainings on L2 learners' acquisition of certain phones across a variety of linguistic environments. Although it is beyond the scope of the present meta-analysis to provide an exhaustive list of all perception training experiments, we have included a few representative studies that cover a broad range of examples. The bulk of training studies have targeted L2 English. Perhaps the single most trained phonemic contrast is the English /ɹ/–/l/ distinction with L1 Japanese participants

https://doi.org/10.1017/S0142716417000418 Published online by Cambridge University Press

(e.g., Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999; Iverson, Hazan, & Bannister, 2005; Lively, Pisoni, Yamada, Tohkura, & Yamada, 1994; McCandliss, Fiez, Protopapas, Conway, & McClelland, 2002). Another popular target for training has been English vowels with L1 Spanish participants (e.g., Aliaga-García & Mora, 2009; Cenoz & García Lecumberri, 1999; Gómez Lacabex, García Lecumberri, & Cooke, 2009; Kondaurova & Francis, 2010). Still others have focused on Chinese speakers' acquisition of English word-final /t/ and /d/ (Flege, 1989) and Korean speakers' acquisition of English codas (Huensch & Tremblay, 2015). Perception training studies have also focused on target languages other than English. For example, at least two studies have targeted English and Japanese speakers' perception of Hindi dental and retroflex stops (Pruitt, 1995; Pruitt, Jenkins, & Strange, 2006). These experiments and countless others investigating different language combinations and target phonemes have successfully trained learners in perceiving target L2 sounds more accurately.

Overall, the studies show that perception training is largely successful, indicating that new phonemic categories have been created in the subjects' mental representations by means of the perceptual modality. If this phonemic system moderates the production modality as well, it is presumed that trained subjects would next be able to show production improvements on the target sounds. However, perception training studies have not been able to reliably show improved production. An oft-cited study by Bradlow, Pisoni, Yamada, and Tohkura (1997) reported that native Japanese speakers' productions of the English /ɹ/ and /l/ improved after perceptual training. Other studies have shown variability in the success of perception training in improving production. Iverson et al. (2012) reported moderate gains in production for English vowels, while Aliaga-García and Mora (2009) showed no improvement in vowel production at all after a 14-week-long perceptual training. A dissertation by García Pérez (2003) demonstrated a similar pattern, that after classroom-based training participants showed perceptual, but not productive improvements. Lopez-Soto and Kewley-Port (2009) reported production gains on some of the 13 codas for which they provided perception training, namely, those sounds for which they also observed large perception gains. Numerous training studies provide conflicting results, making it difficult to ascertain whether or not perception training actually leads to improvement in production.

Despite the inconclusive and conflicting findings from perception training studies measuring production gains, the field of L2 phonology and beyond often cite one popular study by Bradlow et al. (1997) to prematurely proclaim that the perception and production modes are connected because perception training leads to production gains.[1] Researchers should never rely on a single study to answer large and theoretically important questions; one study is insufficient evidence to generalize across all types of target phonemes, language combinations, and training environments, especially when the sample size is quite small (Bradlow et al., 1997, trained 11 participants). Perceptual training's effects on the production modality has been adequately studied in SLA by now to provide a substantive body of research for a meta-analysis to be conducted, which will help bring clarity to the fragmented picture of this domain.

THE PRESENT STUDY

This meta-analysis is a first attempt to compile all perceptual training studies of L2 sounds, conducted across all L1 and L2 combinations that have tested for production gains. By looking at all studies together, SLA researchers will progress in their understanding of the connection between the modalities.

We believe the inconsistent results of perception training's effects on production are in part due to the variability that exists across training program designs. Perception training studies target a vast range of phonemes in various L1 and L2 combinations. In addition, many studies have been conducted across a variety of different research settings (e.g., L2 or foreign language) and have utilized a variety of different training tasks, participant L2 proficiency levels, numbers of trained segments, amounts of speaker variability, and stimulus quality (e.g., naturally recorded or synthesized). We anticipated that these features, in addition to the existence of articulatory and phonetic information, play a role in how effective perception training is on production outcomes. Thus, in addition to calculating effect sizes for all studies, we also coded and analyzed the variables that were expected to have a moderating impact on the outcome of production gains. The research questions that motivated this study are the following:

1. In adult L2 learners, how effective is perception training of L2 phonemes on production outcomes?
2. Is there a relationship between perceptual gains and production gains after perception training, so that the more effective a perception training program is, the greater the transfer effects into the production modality?
3. With regard to manner and place, which phonetic categories show improvements in production after perceptual training?
4. Which features of perception training predict production gains (e.g., training context, length of training, and type of training)?

The fundamental motivation of this study is the theoretical question of the independent or interconnected natures of perception and production. It is therefore necessary to disentangle the two modalities at the point of training. Unfortunately, many training studies do not clearly separate the two modes. For example, many training programs featured in this domain label themselves "perception training," but also include articulatory information and production tasks (e.g., Akita, 2007; Hanlon, 2005). In practical terms, researchers who investigate instructional methods for L2 pronunciation are right to incorporate both modalities: namely, learners with healthy hearing will likely never practice speaking in the complete absence of auditory input, and conversely, they will not likely practice listening without producing a single imitation or utterance. However, in order to test the theoretical point of the perception and production modalities being connected, a thoughtfully designed study should address issues of cross-contamination. At the Interspeech Conference in Tokyo, Hattori and Iverson (2010) called for more studies that train the two modalities in isolation. Therefore, for the purposes of this meta-analysis, a strict operational definition of perception training was adopted, which is summarized graphically in Figure 1. After much discussion, we decided if training regimes prompt the subjects to produce the target sounds, they cannot be

https://doi.org/10.1017/S0142716417000418 Published online by Cambridge University Press

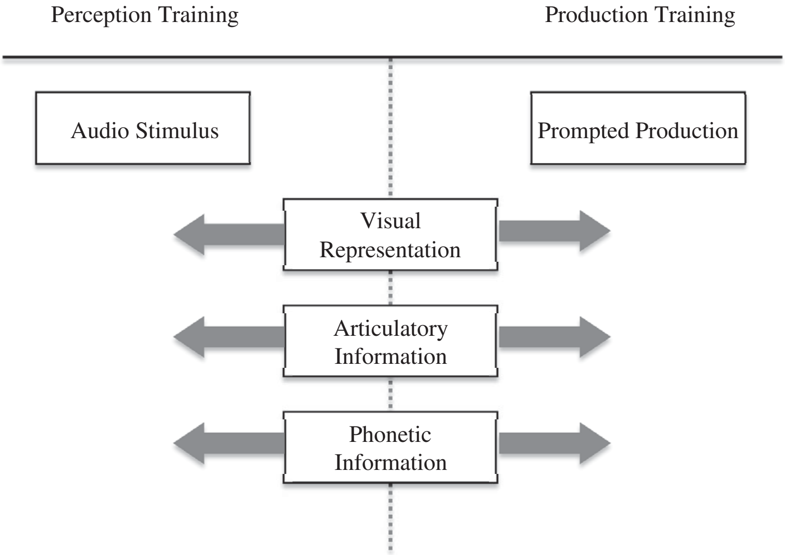

Figure 1. Elements that define perception and production training.

considered pure perceptual training. However, we determined that there are elements that both perception and production training can share, such as providing articulatory information. For example, a learner can be taught that /b/ involves putting both lips together while /v/ uses the upper teeth and lower lip; this articulatory information can aid the learner in perceptually distinguishing between the sounds *or* in producing them correctly. Ultimately, we operationalized a training as perceptually driven if there is a presence of auditory input and the intention to train the perception mode; however, we took a strict stance that perception training cannot involve any sort of prompted production.

METHOD

Literature search

While there are many variations among linguistic perception training experiments, for the purposes of this meta-analysis, we identified six elements that focused our search. We sought to include all empirical investigations that met the following eligibility criteria: studies that (a) were published between 1988 and 2013, (b) were written in English, (c) utilized a training that targeted perception, (d) targeted L2 phonemes, (e) tested participants' production before and after training, and (f) tested adult participants.

We decided that all dissertations, theses, and conference proceedings would be included because they are relatively easy to access through electronic searching, but other types of gray literature (e.g., unpublished reports) were excluded because they were unlikely to be retrieved exhaustively. We considered a problem that affects many domains of social sciences literature, publication bias (Rosenthal, 1979). This situation occurs when experiments do not find significant results and the studies are metaphorically filed away and never shared with colleagues via conferences or journal publications (alternatively, it could be argued that the referees of conferences and journals do not regularly accept these studies for dissemination). To formally assess the potential impact of this "file drawer" problem on our retrieved universe of studies, we utilized a funnel plot (Sterne & Harbord, 2004).

The search for studies began with a comprehensive and exhaustive keyword search of five academic databases (i.e., ERIC, LLBA, MLA, ProQuest Dissertations & Theses, and PyscInfo). The search was limited to studies published from December 31, 1987, to June 30, 2013. More than 5,000 search results were returned. At this point, exact duplicates were eliminated, and any study that contained the following keywords was also eliminated: deaf, cochlear implant, kindergarten, preschool, and agriculture. This resulted in approximately 4,600 potentially relevant studies. After reading every title and abstract for relevance to this meta-analysis, all unrelated studies were discarded, which left a true potential of 633 studies.

In order to obtain the maximum number of potential studies, we then forward-searched the two most frequently cited publications on the topic, Flege (1995a) and Bradlow et al. (1997). This strategy allowed us to uncover approximately 60 more studies. We also hand-searched the table of contents of five relevant journals or publishers: *Journal of Acoustical Society of America*, *Journal of Phonetics*, *Language Learning*, *Phonetica*, and Cambridge Journals (which include *Annual Review of Applied Linguistics*; *Applied Psycholinguistics*; *Bilingualism: Language and Cognition*; *Phonology*; and *Studies in Second Language Acquisition*), which yielded an additional 52 articles. We then checked the references of articles that contained comprehensive reviews on L2 perception and production (e.g., Dziubalska, Wrenbel, & Kul, 2010; Flege, 1995a; Hansen Edwards & Zampini, 2008; Leather, 1999b; Major, 1998; Piske, MacKay, & Flege, 2001; Rauber, Watkins, & Baptists, 2007; Rochet, 1995; Samuel & Kraljic, 2009; Strange, 1995; Tohkura, Vatikiotis-Bateson, & Sagisaka, 1992; Tuller, 2004; Zampini, 1998), special editions of journals (e.g., James, 1994; Leather, 1999a; Major, 1998), and prototypical perception training studies (e.g., Anderson, 2011; Baese-Berk, 2010; Herd, 2011; Iverson et al., 2012; Kissling, 2012). This yielded approximately 50 more studies. Finally, the websites of two prolific researchers in the field, Jim Flege and Anne Bradlow, were searched, and yielded 1 additional study. As the hand searching and footnote chasing began to produce no previously unidentified studies, we believed that our search had become exhaustive. After duplicates were again deleted, we identified a sum of 755 studies as potentially relevant to this meta-analysis, and they were pooled for the beginning of the inclusion and exclusion stage of the project.

Inclusion and exclusion criteria

Seven criteria were identified for determining which studies could provide evidence to answer our research questions and would be included in the meta-analysis. The initial 755 potentially relevant studies were carefully reviewed and filtered through the following inclusion criteria:

1. The study was published between 1988 and 2013. The starting point was chosen because Flege (1988a) served as a watershed moment in the development of the SLM and sparked much research about the interconnected nature of perception and production. The chosen end point reflects the time at which the meta-analysis search was conducted.
2. The report was written in English. Studies reported in languages other than English would ideally be included in this meta-analysis, but the possibility was discounted in order to avoid idiosyncratic language biases, as the researchers found it infeasible to include all languages equally.
3. The study was an empirical investigation of auditory perceptual training or instruction. The training must have included auditory stimuli.
4. The targets of the perceptual training were phonetic segments, rather than suprasegmental or autosegmental features. This decision was primarily made because the SLM addresses segmental categories. In addition, it may be unadvisable to combine segmental, suprasegmental, and autosegmental learning, as the acquisition process underlying these three concepts may be different.
5. The study included quantitative pre- to posttests of participants' production.[2]
6. The target language of the perceptual training was a second or foreign language (FL) for the study participants. The trainings could target any language, and participants could have any L1.
7. The study participants were postpubertal. It has been purported in the critical period hypothesis that there is a developmental window in which it is possible to acquire a language (L1 or L2) to nativelike standards (e.g., Birdsong, 1999; Scovel, 2000; Stölten et al., 2015). However, it is not the aim of this study to test the critical period hypothesis. Rather, in this meta-analysis, studies must have focused on perception training in adults who likely started learning their L2 after puberty, so as to avoid potential confounds in the results by including child data.

One hundred eleven studies were retained after applying the aforementioned inclusion criteria. In the next step, some of the studies that passed the inclusion criteria had to be excluded for one or more of the following reasons:

1. Studies that included a production component in the training that could not be dissociated from the perception component were excluded. If there was evidence that the researcher prompted participants to produce the sounds during the training, thus confounding the two types of trainings, the study was excluded. With this criterion, 53 studies were eliminated (e.g., Akita, 2007; Bettoni-Techio, 2008; García Pérez, 2006; Hanlon, 2005; Lado, 1989; Saito, 2011). Lopez-Soto and Kewley-Port (2009, 2010) were eliminated at this stage as well because there was not enough evidence to be certain the training did not contain prompted production. Studies retained at this point either contained an explicit statement by the researcher that participants were not prompted to produce the sounds during training, or the explanation of the training was so descriptive that we felt confident that participants were not prompted to produce the target sounds.

https://doi.org/10.1017/S0142716417000418 Published online by Cambridge University Press

2. When multiple articles reported on the same data, only one was retained. This case was particularly frequent for dissertations or studies found in conference proceedings that were subsequently published in journals (e.g., Aliaga-García & Mora, 2007, 2009; Bradlow et al., 1997, 1999; Motohashi, 2007, Motohashi-Saigo & Hardison, 2009; Thomson, 2007, 2012; Yeon, 2004, 2008). The study with the most complete data was retained as the representative for inclusion in the meta-analysis, and the alternate reports were utilized to cross-reference additional data as necessary. Twenty-six studies were eliminated for this reason. At this point, 30 studies had passed all inclusion and exclusion criteria. A summary of these studies will be presented in the Results section. However, one goal of the meta-analysis was to calculate effect sizes in order to answer our main research questions. Therefore, one final exclusion criterion was necessary.
3. Studies that did not provide sufficient data (i.e., means, standard deviations, and cell sample sizes on pre- and posttests) to calculate an effect size were eliminated. Before being excluded for this criterion, every attempt was made to contact the author(s) and retrieve necessary data.[3] Twelve studies were unfortunately lost at this point (e.g., Brosseau-Lapré, Rvachew, Clayards, & Dickson, 2013; Handley, Sharples, & Moore, 2009; Rochet, 1995).

The application of the exclusion criteria left a final pool of 18 studies to be investigated in the full meta-analysis.

Coding

The coding scheme for this study was created through an iterative process of coding and revision in a pilot study. The final coding scheme is presented in Table 1. Both researchers coded all studies to ensure that relevant information was not missed, particularly since many of the studies in the meta-analysis were dissertations that often utilized multiple interrelated experiments and samples. The few disagreements that occurred were a result of a misreading of the text, and the discrepancies were discussed and agreed upon.

Effect size calculations

In order to answer the research questions, we calculated Cohen *d* effect sizes for each experimental group. Pre- to posttest (PP) gain scores for a single trained group were calculated by hand using the following equation recommended by Lipsey and Wilson (2001): for the PP effect size calculation ($\mathrm{ES}_{\mathrm{PP}}$),

$$\mathrm{ES}_{\mathrm{PP}} = c_{\mathrm{PP}} \left[\frac{\bar{X}_{\mathrm{post}} - \bar{X}_{\mathrm{pre}}}{SD_{\mathrm{p}}} \right],$$

for the pooled standard deviation (SD_{p}),

$$SD_{\mathrm{p}} = \sqrt{\frac{SD_{\mathrm{pre}}^{2} + SD_{\mathrm{post}}^{2}}{2}}.$$

https://doi.org/10.1017/S0142716417000418 Published online by Cambridge University Press

Table 1. *Coding scheme*

Variables	Values			
Publication Info				
Authors				
Year				
Source	Journal	Book chapter	Dissertation	Conference proceedings
Methodological Features				
Participants' first language				
Target language				
Proficiency	Beginner	Intermediate	Advanced	
Control group	Yes	No		
Sample size				
Mean age				
Substantive Features				
Training				
No. of phones				
Target phones				
Manner				
Place of articulation				
Stimuli	Natural	Synthetic		
No. of speakers				
Context	Classroom	Laboratory	At home	Mixed
Setting	Second language	Foreign language	Mixed	
Mode of delivery	Face-to-face	Technology mediated		
Group size	Individual	Small group	Large group	
Type of task	Identification	Discrimination	Other	
Variability	Yes	No		
Length				
Total hours				
No. of sessions				
Length of session (min)				
Time period (days)				
Tokens per session				
Feedback	Yes	No		
Phonetic instruction	Yes	No		
Articulatory information	Yes	No		

Table 1 (*cont.*)

Variables	Values			
Visual representation	Yes	No		
Orthographic representation	Yes	No		
DV measure, perception	Identification	Discrimination	Other	
DV measure, production	Free	Controlled		
Elicitation prompt	Auditory	Orthographic	Both	
Length of speech sample	Single word	Sentence	Passage	Mixed
Analysis	Human rater	Acoustic		

A second type of effect size was calculated by hand for the studies that included both experimental and control group data for pre- and posttests (pre- to posttest with control; PPC). We decided that it was important to include the PPC effect size calculation in addition to the single trained group PP effect size because it reports the experimental group performance in light of the control group's performance all in one score. Morris (2008) recommends the following equation when calculating this type of effect size (T is the trained group, C is the control group): for the PPC effect size calculation ($\mathrm{ES}_{\mathrm{PPC}}$),

$$\mathrm{ES}_{\mathrm{PPC}} = c_{\mathrm{PPC}} \left[\frac{\left(M_{\mathrm{post},T} - M_{\mathrm{pre},T}\right) - \left(M_{\mathrm{post},C} - M_{\mathrm{pre},C}\right)}{SD_{\mathrm{pre}}} \right],$$

for the pooled standard deviation at pretest (SD_{pre}),[4]

$$SD_{\mathrm{pre}} = \frac{\sqrt{(n_T - 1)}\, SD^2_{\mathrm{pre},T} + (n_C - 1)\, SD^2_{\mathrm{pre},C}}{n_T + n_C - 2}.$$

In both the PP and PPC calculations, we included a correction for sampling bias (c), as recommended by Morris (2008):

$$c_{\mathrm{PP}} = 1 - \left[\frac{3}{4(n-1) - 1} \right],$$

$$c_{\mathrm{PPC}} = 1 - \left[\frac{3}{4(n_T + n_C - 2) - 1} \right].$$

In order to interpret the within-group effect sizes, we used the standards for SLA that were established by Plonsky and Oswald (2014), who rightly pointed out that the benchmarks offered by Cohen (1988) were never intended to provide a standard for all fields of behavioral sciences research. After reviewing the distribution of effect sizes in 346 primary studies and 91 meta-analyses in the field of L2 studies, Plonsky and Oswald (2014) offered the following field-specific benchmarks for within-group, pre- to posttest effect sizes such as ours: small = 0.6, medium = 1.0, and large = 1.4. To interpret the correlation coefficient, Plonsky and Oswald (2014) suggest that small = 0.25, medium = 0.40, and large = 0.60. We also present the production effect sizes from the present study in light of the L2 pronunciation meta-analysis conducted by Lee, Jang, and Plonsky (2015).

Several studies in this meta-analysis utilized multiple tests to measure perception and production gains. For example, Hazan et al. (2005) employed both a native speaker rating and a native speaker identification task to measure participants' productions. When one study reported multiple dependent measures, one test was selected to represent each modality. For the perception pre- and posttests, the identification task was by far the most common testing measure; therefore, if there were multiple perception tests, the identification task was chosen as the representative test. If there were multiple identification tests, the one that was closest to the training task in stimuli and phonetic context was chosen. For the production test, elicited production with native speakers as judges in an identification task was the most frequently used measure. Therefore, if multiple production measures or analyses existed in one study, the elicited production with native speaker identifications was selected as the representative task. We chose to ensure independence of effect size scores in this way because of the importance of considering how the dependent measures generalize from the trained tasks. (The topic of generalization is discussed in more detail in the Discussion section of this article.) However, in order to present how substantive features in the dependent measure may influence the magnitude of effect of the training, we have also included a separate analysis of effect sizes from all production dependent measures that each study reported. These data are presented along with the first research question.

RESULTS

Thirty studies passed all inclusion and exclusion criteria with the exception of reporting sufficient data to calculate effect sizes. Before answering the research questions, we will present a synthesis of the 30 studies here. Table 2 displays all of the studies and basic information about each one (e.g., L1, L2, target phoneme, and number of experimental groups). Some of the publications included multiple, independent experimental groups; for example, Hardison (2003) reported on 6 different experimental groups. Therefore, the 30 reports produced 51 unique experimental groups that had an average sample size of 11.61. One author did not report findings for production gains after perception training on 2 experimental groups, but of the 49 experimental groups that did report this information, 36, or 73.47%, reported positive gains on at least one production measure after perception training.

These 30 reports were published in dissertations ($k = 12$), journals ($k = 13$), conference proceedings ($k = 4$),[5] and book chapters ($k = 1$). They represent a wide

https://doi.org/10.1017/S0142716417000418 Published online by Cambridge University Press

Table 2. *Thirty studies that passed inclusion and exclusion criteria*

Study	L1	L2	Segments	No. of Experimental Groups *k*	*n*	Groups With Production Gains
Anderson (2011)	English	Spanish	ɾ, r	1	21	1
Baese-Berk (2010)	English	Artificial	d, t	2	18, 14	1
Bradlow et al. (1997)	Portuguese	English	ɹ, l	1	11	1
Brosseau-Lapre et al. (2013)	English	French	ə, ø	6	10 each	0
Counselman (2010)	English	Spanish	e, o	1	13	1
Gómez Lacabex & García Lecumberri (2010)	Spanish	English	ə	1	17	1
Hamada & Tsushima (2001)	Japanese	English	s, θ, z, ð, b, v, l, ɹ, 4V	2	14, 11	2
Han (2002)	Korean	English	ɹ, l	1	18	1
Handley et al. (2009)	Mandarin	English	ɹ, l	2	6, 6	UR
Hardison (2003)	Japanese	English	ɹ, l	6	8, 8, 2, 2, 2, 2	6
Haslam (2011)	Japanese	English	ɹ, l	2	5, 6	0
Hazan et al. (2005)	Japanese	English	ɹ, l	3	10, 10, 5	3
Herd et al. (2013)	English	Spanish	ɾ, r, d	1	10	1
Huensch (2013)	Korean	English	ʃ, tʃ, dʒ	1	12	1
Imaizumi et al. (1998)	Japanese	English	ɹ, l	1	10	1
Iverson et al. (2012)	French	English	14V	2	15, 21	2
Lambacher et al. (2005)	Japanese	English	æ, ɑ, ʌ, ɔ, ɝ	1	20	1
Lengeris (2009)	Greek	English	i, ɪ, e, æ, ʌ, ɑ, ɜ, ɒ, ɔ, u	1	18	1

https://doi.org/10.1017/S0142716417000418 Published online by Cambridge University Press

Table 2 (*cont.*)

				No. of		
				Experimental		Groups With
				Groups		Production
Study	L1	L2	Segments	*k*	*n*	Gains
Motohashi (2007)	English	Japanese	ss, tt, kk	2	15, 15	2
Nobre-Oliveira (2007)	Portuguese	English	i, ɪ, ɛ, æ, ʊ, u	2	10, 13	1
Reis & Nobre-Oliveira (2007)	Portuguese	English	p, t, k	1	11	1
Rochet (1995)	Mandarin	French	p, b	1	12	1
Sawallis & Townley (2009)	Japanese, Korean, Chinese	English	ɹ, l	1	UR	1
Soler-Urzua (2011)	Spanish	English	ɪ	2	17, 16	1
Stenning & Jamieson (2002)	Spanish	English	i, ɪ, e, ɛ, æ	1	UR	1
Thomson (2007)	Mandarin	English	i, ɪ, e, ɛ, æ, ɒ, ʌ, o, ʊ, u	2	11, 11	2
Todaka (2008)	Japanese	English	ɹ, l	1	6	1
Underbakke (1993)	Japanese, Thai	English	ɹ, l	1	26	1
Wang (2002)	Mandarin, Cantonese	English	i, ɪ, ɛ, æ, ʊ, u	1	16	0
Yeon (2004)	Korean	English	ʃ, tʃ, dʒ	1	15	0

Note: L1, first language; L2, second language; UR, unreported; V, vowels.

variety of L1s, with Japanese being the most represented language and English following as a close second: Japanese ($k = 7$), English ($k = 6$), Portuguese ($k = 3$), Spanish ($k = 3$), Korean ($k = 3$), Mandarin ($k = 3$), French ($k = 1$), Greek ($k = 1$), and mixed L1s ($k = 3$). Overwhelmingly, the target L2 was English: English ($k = 23$), Spanish ($k = 3$), French ($k = 2$), Japanese ($k = 1$), and an unspecified language ($k = 1$).[6] Many of the studies reported a description of participants' ages (e.g., undergraduates), but of those that reported a mean, the average age of participants was 24.85 ($SD = 5.5$). Of the 15 studies that described the sample population's L2 proficiency, 5 reported that learners were novices or beginners, 9 studies' participants were intermediate learners, and only 1 reported on high-level learners. Sixteen of the studies included a control group, and 7 administered a delayed posttest.

This sample of 30 studies employed a variety of training programs situated in various contexts. The following describes the training programs of those that reported this information in detail. Sixteen studies reported that training occurred in an FL context, 11 were in an L2 context, and two had groups in both FL and L2 environments. Fourteen studies reported that participants completed the perception training in a laboratory, 4 completed it at home, 2 in a classroom, and 2 in more than one setting (e.g., in the classroom and at home). Eighteen of the studies targeted consonants during training, 11 targeted vowels, and 1 focused on a combination of consonants and vowels. The average number of phoneme targets was 3.83 ($SD = 3.36$, mode = 2). One-third of the studies ($k = 10$) consisted of training the English /ɹ/–/l/ distinction. Of the studies that reported definitive numbers, the average number of hours spent training was 4.79 ($SD = 4.64$),[7] in 7.7 training sessions ($SD = 8.5$), over the course of 19.54 days ($SD = 15.72$).

The most common type of training was an identification task, where participants heard an auditory stimulus and pushed a button that corresponded to the sound they thought they heard ($k = 18$). All of the studies that utilized an identification task gave participants feedback. Although one study did not report this information, the remaining 17 that utilized an identification task used variability of some sort (e.g., speaker, phonetic context, and feature variation). Three identification tasks involved a fading technique (i.e., Imaizumi, Itoh, Tamekawa, Deguchi, & Mori, 1998; Rochet, 1995; Wang, 2002), in which participants were first trained on the most extreme or enhanced version of the target phonemes and eventually moved to less extreme variations of the targets that appeared closer together on some sort of continuum; and two identification training programs used an adaptive technique (i.e., Iverson et al., 2012; Lengeris, 2009), which means that participants received more practice with trials that were difficult for them.

Experimental groups in eight studies underwent a more unique type of training. Nobre-Oliveira (2007), Reis and Nobre-Oliveira (2007), and Underbakke (1993) had participants do a combination of identification and discrimination tasks; Handley et al. (2009) used an oddball discrimination training program; participants in Counselman (2010) completed a listening homework task; the learners in Gómez Lacabex and García Lecumberri (2010) listened to audio stimuli and induced rules; Baese-Berk (2010) used a passive listening association training; and Soler-Urzua (2011) used a text to speech activity.

https://doi.org/10.1017/S0142716417000418 Published online by Cambridge University Press

Table 3. *Pre- to posttest effect size scores for 18 studies*

Study	Independent Experimental Groups	Perception Effect Size	Production Effect Size
Anderson (2011)		0.42	0.30
Bradlow et al. (1997)		1.27	0.37
Counselman (2010)		—	0.59
Gomez Lacabex & Garcia Lecumberri (2010)		—	0.71
Han (2002)		—	0.84
Hazan et al. (2005)	Audio only	1.38	0.09
	Audiovisual (natural)	1.07	0.38
	Audiovisual (synthetic)	1.10	0.13
Herd et al. (2013)		0.47	0.45
Huensch (2013)		0.96	1.22
Lambacher et al. (2005)		0.46	0.41
Lengeris (2009)		−0.03	−0.08
Motohashi (2007)	Audio only	0.69	1.24
	Audiovisual	1.77	1.65
Nobre-Oliveira (2007)	Natural tokens	1.23	−0.03
	Synthesized tokens	2.05	1.25
Reis & Nobre-Oliveira (2007)		−1.97	0.78
Soler-Urzua (2011)	Text to speech	0.48	0.15
	Nontext to speech	−0.09	0.04
Thomson (2007)	Long vowel training	0.65	0.50
	Select vowel training	1.26	0.51
Underbakke (1993)		1.35	0.32
Wang (2002)		2.69	0.72
Yeon (2004)		2.11	0.46
Mean (*SD*)		0.92 (0.96)	0.54 (0.45)

Note: Counselman (2010), Gomez Lacabex and Garcia Lecumberri (2010), and Han (2002) did not administer a pre- and posttest of perception.

Research question 1 (RQ1): In adult L2 learners, how effective is perception training of L2 sounds on production outcomes?

Eighteen studies passed all inclusion and exclusion criteria and included sufficient data to calculate effect sizes. In order to address the first research question, the Cohen *d* was calculated for 21 independent (or unique) experimental groups for perception scores, and 24 independent experimental groups for production scores. Table 3 presents the 18 studies, their unique experimental groups, and all of the individual perception and production effect sizes. The average PP effect size for perception outcomes after perception training was 0.92, $SD = 0.96$, $k = 21$, confidence interval (CI) [0.51, 1.33], and the average PP effect size of production outcomes after perception training was 0.54, $SD = 0.45$, $k = 24$,

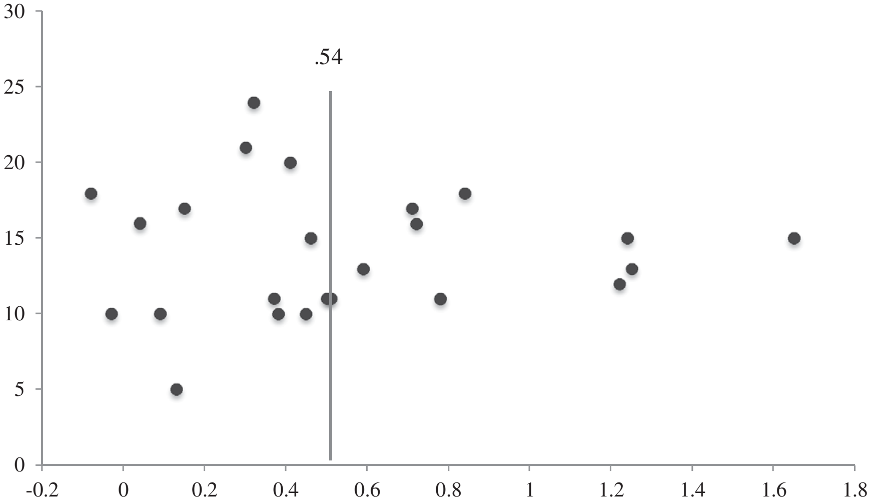

Figure 2. Funnel plot of production effect size by sample size.

CI [0.36, 0.72]. Thus, the means suggest perception training affords medium-sized gains on perception, and the experimental groups experience a small but trustworthy improvement in their production after perception training (since the mean effect size confidence interval does not contain or touch zero). That is, perception-only training does carry over to the production modality with a small, yet robust impact. The meta-analysis conducted by Lee et al. (2015) found that effects of traditional pronunciation instruction was $d = 0.83$. Comparing this to the present study's findings, it is clear and not surprising that perception-only training leads to smaller gains in production than more traditional pronunciation instruction, which presumably includes production practice. It is important to note that the standard deviations for both the perception and production effect size calculations in the present study are almost as large or larger than the means, which indicates great variability of effects across studies, possibly due to differences in study design and other moderating variables.

A funnel plot was used to determine if publication bias exists in this domain of literature. The funnel plot is able to show if there is a missing representation of results that are small to nonsignificant (Sterne & Harbord, 2004). Figure 2 displays the funnel plot of effect sizes of production outcomes. In training studies, there is not likely to be adverse outcomes represented in negative values. Accordingly, the plot shows a general funnel shape without missing data points, which indicates there is not a likely publication bias in this type of literature.

Approximately 60% of the 18 studies included a control group. From these studies, we were able to calculate 10 unique effect sizes for perception outcomes, and 11 unique effect sizes for production outcomes using the PPC effect size calculation, which incorporates both the experimental and the control groups,' pre- and posttest data into one equation. Including data from the control group

https://doi.org/10.1017/S0142716417000418 Published online by Cambridge University Press

helps to account for any extraexperimental exposure to the target phonemes and/or test–retest effects. Overall, effect sizes changed slightly in both directions compared to the PP calculation of the experimental group data only. The mean PP effect size for perception outcomes of the 10 unique groups for the experimental condition only was 1.12 (*SD* = 0.87); when we factor in the control groups (PPC), the effect size drops slightly to 0.93 (*SD* = 0.72). This small decrease in effect is not unexpected when within-group effect sizes are compared to between-group contrasts, as participants serve as their own controls which decreases variance (Plonsky & Oswald, 2014). Perception training has a medium-sized effect on perception outcomes whether experimental group data are analyzed alone or in combination with control group data. As the mean effect size of training remained approximately the same for perception outcomes with or without control groups, the positive effects for perception after training were unlikely due to extraexperimental exposure or test–retest effects.

For production outcomes, the mean PP effect size for experimental groups only was 0.61 (*SD* = 0.55); when we factor in the control groups, the effect size (PPC) increases to 0.89 (*SD* = 0.61; see Table 4). Including the control group scores increases the effect of perception training on production outcomes from a small effect to nearly a medium effect. Factoring in the control groups' performance affords an enhancement in the interpretation of the effects of perception training on production outcomes. In other words, the trained groups' improved performance on production tasks after perception training appears even more substantial in light of the control groups' lower performance in the posttest session compared to the pretest. Having a control group makes an experimental design more robust, and calculating the PPC effect size is better than the PP effect size of an experimental group alone. However, in terms of effect size calculation for this meta-analysis, the number of unique experimental groups contributing to the mean is higher for the PP calculation (*k* = 24) than the PPC calculation (*k* = 11). It would be ideal to utilize the PPC calculation with a larger *k*. Thus, with these data, one should be cautious of prioritizing the PPC calculation because of the small sample. Nevertheless, we can conclude that the effects of perception-only training on production gains are present, whether or not control groups are included in the calculations.

Although we chose to handle dependencies by selecting one representative test per modality per unique participant group, we also recognize the value of meta-analyzing all dependent measures to paint a better picture of how a particular task can affect the appearance of gains after training. The 24 unique experimental groups were tested on an average of 1.79 production tests, with the range being 1 to 6 tests per group. The average effect size of all production tests after perception training was *d* = 0.58, *SD* = 0.72, *k* = 43, CI [0.36, 0.79], which is remarkably similar to the PP effect size of the one representative production test (*d* = 0.54). Table 5 lists three substantive features of the dependent measures (i.e., elicitation method, speech response length, and type of analysis) and their corresponding effect sizes. All dependent measures were coded as controlled in nature as opposed to free speech; thus, this category could not be presented in the table. Orthographic elicitation prompts encouraged slightly larger effects, *d* = 0.60, *SD* = 0.79, *k* = 34, CI [0.33, 0.87], than auditory prompts, *d* = 0.47, *SD* = 0.12, *k* = 4,

https://doi.org/10.1017/S0142716417000418 Published online by Cambridge University Press

Table 4. *Pre- to posttest (PP) effect size scores presented alongside PP with control effect size scores*

Study	Independent Experimental Groups	PP Perception Effect Size		PP Production Effect Size	
		Experimental Only	With Control	Experimental Only	With Control
Bradlow et al. (1997)		1.27	2.21	—	—
Herd et al. (2013)		0.47	0.72	0.45	1.00
Lambacher et al. (2005)		0.46	0.30	0.41	0.44
Motohashi (2007)	Audio only	0.69	0.36	1.24	1.22
	Audiovisual	1.77	1.30	1.65	1.61
Nobre-Oliveira (2007)	Natural tokens	—	—	−0.03	0.89
	Synthesized tokens	—	—	1.25	2.19
Soler-Urzua (2011)	Text to speech	0.48	0.53	0.15	0.34
	Nontext to speech	−0.09	0.04	0.04	0.22
Underbakke (1993)		1.35	0.64	0.32	0.65
Wang (2002)		2.69	1.92	0.72	0.95
Yeon (2004)		2.11	1.28	0.46	0.31
Mean (*SD*)		1.12 (0.87)	0.93 (0.72)	0.61 (0.55)	0.89 (0.61)

https://doi.org/10.1017/S0142716417000418 Published online by Cambridge University Press

Table 5. *Pre- to posttest effect sizes of all production dependent measures presented by testing feature*

				Confidence Interval	
Production Test	*k*	Mean (*SD*)	*SE*	Lower Bound	Upper Bound
Elicitation prompt					
Auditory	4	0.47 (0.12)	0.06	0.35	0.59
Orthographic	34	0.60 (0.79)	0.14	0.33	0.87
Both	5	0.48 (0.18)	0.08	0.33	0.64
Output length					
Word	27	0.34 (0.33)	0.06	0.22	0.47
Sentence	6	0.66 (0.18)	0.07	0.52	0.81
Passage	6	1.24 (1.56)	0.64	−0.01	2.48
Multiple	4	1.03 (0.73)	0.36	0.31	1.74
Analysis					
Human rater	29	0.50 (0.45)	0.08	0.33	0.66
Acoustic	14	0.74 (1.09)	0.29	0.17	1.31

CI [0.34, 0.59]. Single word elicitations were by far the most commonly requested speech response length (k = 27), but the effect sizes were larger for responses at the passage length, d = 1.24, SD = 1.56, k = 6, CI [–0.01, 2.48], and mixed lengths, d = 1.03, SD = 0.73, k = 4, CI [0.31, 1.74]. Finally, analyses that acoustically measured participants' speech showed larger effects, d = 0.74, SD = 1.09, k = 14, CI [0.17, 1.31], than analyses utilizing human raters, d = 0.50, SD = 0.45, k = 29, CI [0.33, 0.66]. These data make it evident that the characteristics of the dependent measure can impact the magnitude of gains after training.

RQ2: Is there a relationship between perceptual gains and production gains after perception training?

In order to answer this research question, we used the PP effect sizes of the experimental groups alone for perception and production data and correlated the two. Three studies were not included in this analysis because perception data were not reported (i.e., Counselman, 2010; Gómez Lacabex & García Lecumberri, 2010; Han, 2002), which left a sample of 21 unique experimental groups to be analyzed.[8]

Perception and production PP effect sizes for the 21 unique experimental groups were graphed on a histogram to test for normal distribution. The data were not evenly distributed: the perception data were skewed to the left, and the production data were skewed to the right. Thus, a nonparametric correlation was run. No statistically significant correlation was found between the perception and production data, although the correlation coefficient is evidence of a small to medium relationship between the two modalities (r = .31, p = .18). Figure 3 depicts the data on a scatterplot graph, and the visual representation also suggests a positive relationship between the two modes. It seems intuitive that the more effective a perception training program is, the greater the transfer effects

https://doi.org/10.1017/S0142716417000418 Published online by Cambridge University Press

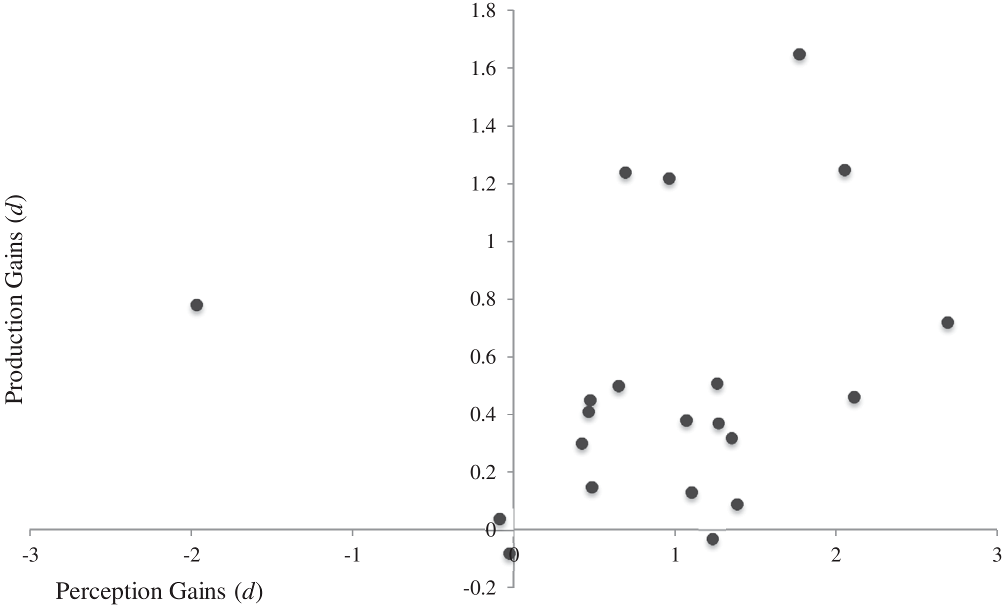

Figure 3. Scatterplot of perception and production gains.

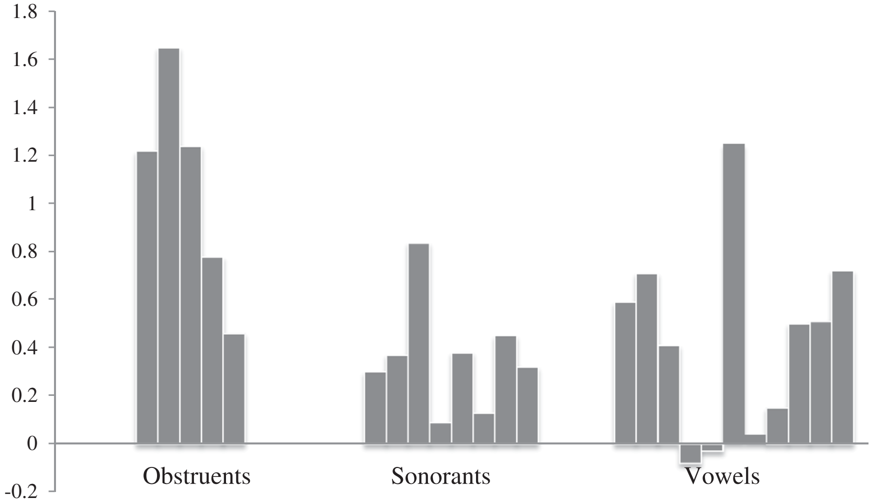

Figure 4. Effect size (pretest to posttest) by manner of articulation.

into the production modality. The nonsignificant findings in the statistical analysis may be due to the small number of experimental groups ($k = 21$), and a more statistically robust relationship may have been found with more data. However, with the results at hand, it seems that the set of factors that stimulate perception improvements are not entirely independent of those that stimulate production improvements.

RQ3: With regard to manner and place, which phonetic categories show improvements in production after perceptual training?

In terms of manner, the target phonemes were collapsed into three categories: vowels, sonorants, and obstruents, and PP effect sizes were used to calculate means for each category. Production of vowels, $d = 0.43$, $SD = 0.40$, $k = 11$, CI [0.20, 0.67], and sonorants, $d = 0.36$, $SD = 0.23$, $k = 8$, CI [0.20, 0.52], improved to a similar small degree after perception training. In contrast, obstruents improved to a larger degree, $d = 1.07$, $SD = 0.46$, $k = 5$, CI [0.67, 1.47].

During the analysis, it became apparent that the there was an obstacle that prevented the data from being analyzed in terms of place of articulation (see Figure 4 and Table 6). Ten out of 18 studies targeted phonemes of different places of articulation within the same training. For example, Reis and Nobre-Oliveira (2007) trained participants on voice onset times (VOTs) of the bilabial, alveolar, and velar stops, /p/, /t/, and /k/. The pre- and posttest scores of this particular study are collapsed, encompassing all three phonemes, and it was not possible to separate the data to capture the gains of the different places of articulation. Thus, we determined that this type of categorical moderator analysis was not possible for the current meta-analysis.

https://doi.org/10.1017/S0142716417000418 Published online by Cambridge University Press

Table 6. *Effect sizes by manner of articulation*

				Confidence Interval	
Target Segment	*k*	Mean (*SD*)	*SE*	Lower Bound	Upper Bound
Manner					
Vowel	11	0.43 (0.40)	0.12	0.20	0.67
Sonorant	8	0.36 (0.23)	0.08	0.20	0.52
Obstruent	5	1.07 (0.46)	0.21	0.67	1.47

RQ4: Which features of perception training predict production gains?

In order to answer the last research question, PP effect sizes were grouped according to substantive features of the training. Table 7 presents the number of studies that fall into each category, the average effect size, and the 95% CI for each type of substantive feature.

Table 7 should be interpreted keeping in mind that it reveals substantive features of the training programs that may have had the largest impact on improving production after perception-only training. It does not describe the perception training features that influence the greatest gains in perception. A separate meta-analysis on all perception training studies would be able to reveal the features of training programs that most effectively improve perception.

Participants improved production to a greater extent in the L2 setting than in the FL or mixed setting. In addition, production improved to a greater degree for beginners than for intermediate learners. Both of these moderating variables reflect the findings of the meta-analysis on pronunciation instruction (Lee et al., 2015). The type of training task did not seem to have a large impact on effect sizes: identification tasks, identification and discrimination tasks, and other types of tasks all induced a small magnitude of production gains. However, the location of the training encouraged greater gains when completed at home versus in the laboratory. Training programs that lasted less than 3 hr were more effective than longer training programs. This seems counterintuitive, and actually contradicts the findings of Lee et al. (2015), which showed that longer pronunciation treatments are more effective than shorter ones. It does, however, resonate with the trend reported by Norris and Ortega (2000) for overall L2 instruction studies lasting 2 hr or less being more effective than longer treatments. The total number of sessions did not seem to have an effect on the degree of production change. L2 instructors, learners, and researchers may be encouraged to know that production gains can occur in fewer and shorter training programs.

In terms of the number of trained phonemes, training programs that targeted one or two phonemes only improved by a small magnitude, $d = 0.36$, $SD = 0.26$, $k = 11$, CI [0.20, 0.51], while training three phonemes seemed to be the most effective, $d = 0.97$, $SD = 0.48$, $k = 6$, CI [0.58, 1.35], and training four or more phonemes also improved by a small magnitude, $d = 0.47$, $SD = 0.45$, $k = 7$, CI [0.13, 0.80]. Thus, the data seem to support that slightly larger sets of trained phonemes improve

Table 7. *Substantive features of the perception-only training and the pre- to posttest effect size calculations of production gains*

				Confidence Interval	
Substantive Feature	*k*	Mean (*SD*)	*SE*	Lower Bound	Upper Bound
Setting					
SL	8	0.83 (0.48)	0.17	0.49	1.16
FL	13	0.45 (0.39)	0.11	0.23	0.65
SL & FL	3	0.20 (0.16)	0.09	0.02	0.38
Proficiency					
Beginner	5	0.84 (0.58)	0.26	0.33	1.35
Intermediate	11	0.46 (0.46)	0.14	0.19	0.73
Training Context					
Laboratory	13	0.35 (0.20)	0.05	0.24	0.45
At home	5	0.92 (0.68)	0.30	0.33	1.52
Type of training task					
Identification	15	0.56 (0.47)	0.12	0.32	0.80
Identification & discrimination	4	0.58 (0.56)	0.28	0.03	1.13
Other	5	0.47 (0.35)	0.16	0.16	0.77
Total length of training					
Short, <3.5 hr	13	0.68 (0.46)	0.13	0.43	0.93
Long, ≥3.5 hr	7	0.24 (0.19)	0.07	0.09	0.38
No. of sessions					
Short, <6	11	0.48 (0.42)	0.13	0.23	0.73
Long, ≥6	13	0.60 (0.48)	0.13	0.34	0.86
No. of trained phonemes					
1–2	11	0.36 (0.26)	0.08	0.20	0.51
3	6	0.97 (0.48)	0.20	0.58	1.35
≥4	7	0.47 (0.45)	0.17	0.13	0.80
Stimuli					
Natural	16	0.48 (0.49)	0.12	0.24	0.72
Synthesized	5	0.60 (0.43)	0.19	0.22	0.98
Speaker variability					
Yes	17	0.45 (0.38)	0.09	0.27	0.63
No	4	0.81 (0.76)	0.38	0.07	1.56
Phonetic instruction					
Yes	5	0.99 (0.64)	0.29	0.43	1.55
No	17	0.40 (0.31)	0.08	0.25	0.55
Articulatory information					
Yes	5	0.51 (0.53)	0.24	0.05	0.98
No	17	0.54 (0.46)	0.11	0.32	0.76
Orthographic representation					
Yes	18	0.52 (0.50)	0.12	0.29	0.75
No	4	0.52 (0.20)	0.10	0.33	0.72

the most, lending weight to Nishi and Kewley-Port's (2007) encouragement to train larger sets of phonemes at the same time if possible. However, five out of six of the studies that trained three phonemes focused on obstruents, which suggests that this category of multiple phoneme targets may be inflated.

As far as the training stimuli, it seemed to not matter whether training programs used natural or synthesized tokens, although synthesized tokens encouraged a slightly larger impact. Variability did not seem to be a necessary component of a successful training program. Studies that did not include speaker variability led to a larger effect of $d = 0.81$, $SD = 0.76$, $k = 4$, CI [0.07, 1.56], compared to studies that did include speaker variability, $d = 0.45$, $SD = 0.38$, $k = 17$, CI [0.27, 0.63]. However, two of the four studies that did not include speaker variability were studies that targeted obstruents. Again, the obstruent data may be driving up the mean score of particular groupings of studies.

Production outcomes after trainings that included phonetic instruction were much higher than studies with no phonetic instruction, $d = 0.99$, $SD = 0.64$, $k = 5$, CI [0.43, 1.55], and $d = 0.40$, $SD = 0.31$, $k = 17$, CI [0.25, 0.55], respectively. In contrast, any type of articulatory information or orthographic representation of the trained phonemes did not seem to impact the degree of pronunciation improvement.

When interpreting the information in Table 6, it is important to note the *k* values. Categories that were only represented by one effect size were not included to guard against statistical misrepresentation. In addition, the important information conveyed through confidence intervals in this table is that all of the observations are statistically robust as none of the lower boundaries touches or includes zero.

DISCUSSION

The results of the present meta-analysis suggest that (a) perception-only training leads to small-sized production gains; (b) there is a small to medium, not statistically significant relationship between perception gains and production gains; (c) production gains are larger for obstruents than for vowels or sonorants; and (d) there are five features of perception-only training that likely encourage larger production gains: second language contexts, beginner-level of L2 experience, training at home, a short training, and the existence of phonetic instruction. These results have both theoretical and methodological importance.

The answer to the first research question, that perception-only training leads to production gains, indicates that the two modalities are connected, insomuch as training in the perception mode can induce positive change in production. From a language acquisition perspective, these results are encouraging and theoretically informative. Proponents of skill acquisition theory believe that learned knowledge is "so highly specific that it does not transfer well, even to what may seem quite similar tasks" (DeKeyser, 2015, p. 97). This theory states that practice in one area will improve that skill, and it is not likely that gains will transfer to other related skills. For example, comprehending an L2 does not automatically transfer to speaking that language. However, the results of this meta-analysis show that one training in sound perception is advantageous for two modalities, as perception and production both improve. These results perhaps complicate skill acquisition

https://doi.org/10.1017/S0142716417000418 Published online by Cambridge University Press

theory, but are also reassuring to L2 teachers and learners. Phonetic instruction can be efficient, as time spent on listening practice can benefit pronunciation as well.

The results from all four research questions taken together contribute to the theoretical discussion of the nature of the connection between the perception and production modes during L2 phonological acquisition. Recall that the SLM posits that the perception and production modes are connected via the mental representation. This meta-analysis offers a concrete example of this process in action. Theoretically, perception training informs the mental representation to become more targetlike, which leads to improvements in production. However, the mental representation of a phoneme contains both perceptual and articulatory information, which creates an environment for two possibilities: a more accurate perceptual representation leads to more accurate production, or an element of the perception training actually modifies the productive information in the mental representation, which in turn affects the production mode. Results from the fourth research question show that the existence of phonetic instruction encouraged greater production gains. This suggests that the perception mode could have been side stepped with this component of the procedure. For example, Han (2002) taught Korean learners the English /ɹ/–/l/ distinction, with the first training session consisting of a 50-min lesson about the phonetic and articulatory characteristics of Korean and English liquids. This information could have directly fed the productive information in the mental representations of the phonemes. Another possibility is that the phonetic instruction encouraged participants to attend to articulatory information in the auditory signal during the perception training. It is possible that the factors that affect production improvements may be fully or partially independent of perception.

In sum, the meta-analysis shows that perception and production are connected in L2 development, but it does very little to provide evidence for any other postulates of the SLM. We cannot say that perception must precede production, and we cannot definitely state that as perception accuracy increases, production accuracy increases. We can confirm that there is a transfer of improvement in the direction tested here, from perceptual mode to production mode, but the perception training paradigm cannot offer evidence of a bidirectional relationship where production-only training influences the perception mode. A more detailed model of L2 speech learning will include distinctions in manner of articulation, the bi- or unidirectional transfer of influence, in addition to answering the questions set forth by the current hypotheses of Flege's SLM, namely, whether production can precede perception.

It is important to look at the results of this meta-analysis with a critical eye in light of two facts: the average training took place over the course of 20 days, and the L2 setting encouraged larger gains than the FL environment. The first point alone is sufficient to demonstrate that participants had the opportunity to interact with the target phones outside of the training. Control group data indicates that extra-experimental exposure did not effect positive gains in production. However, the fact that the L2 context encourages larger effects than FL environments shows that the environment did play a role. It is possible that the type of participant living in the L2 environment is different than those who live in the FL environment. It could also be that the environment *in conjunction with* undergoing a perception-only training encourages participants to improve pronunciation. Participants undergoing training may have increased motivation to talk to family and friends about the target

https://doi.org/10.1017/S0142716417000418 Published online by Cambridge University Press

phonemes, or it may lead to a heightened awareness of the sounds when interacting in the L2 environment. These speculations support an explanation that off-task, long-term interaction with phonemic information, not necessarily limited to only one modality, can induce change in production.

The ability to interact in a long-term setting opens an important discussion about online and long-term processing of stimuli. Cognitive neuroscience (e.g., Campbell et al., 2001; Wilson et al., 2004) shows that when humans hear a speech sound, the productive areas of the brain are simultaneously activated in immediate online processing, and vice versa, the auditory areas of the brain are activated when viewing silent videos of mouthed speech. In contrast, the L2 training paradigm is fundamentally different. That training programs last longer than minutes, that they last days, weeks, or months, provides participants time to interact with the target phonemes in which any modality can be activated. Any number of conscious, explicit thoughts and discussions can occur, resulting in changes in either modality. Thus, the findings of this meta-analysis are valuable in terms of L2 teaching, acquisition, and development, but one should be cautious of using L2 training experiments as strong evidence for the cognitive question of a perception–production connection. In the future, L2 training methodology can contribute to the cognitive question if researchers isolate the perception training program to one session and do not allow participants to speak or mouth sounds during the training. There are very few studies that have attempted to train perception or production this fast (e.g., Haslam, 2011; Kartushina, Hervais-Adelman, Frauenfelder, & Golestani, 2015; Reis & Nobre-Oliveira, 2007), and this is an important design for future research. The L2 phonetic training paradigm has the ability to speak to both the long-term entangled relationship of perception and production like the L1 and deaf and hard of hearing literature, and the simultaneous or more immediate processing of the modalities in the brain like cognitive neuroscience. At this point, most of the L2 training literature has focused on the former, but researchers can certainly design studies that contribute information to the latter.

The third research question revealed that obstruents trained better than sonorants or vowels. One explanation for this result could be that obstruent articulation is easier to hear, analyze, and subsequently mimic than, for example, the slight changes in the tongue body necessary to produce a vowel. The specific articulation involved in obstruent production could somehow be more salient (aurally or visually) than sonorants or vowels, and the mirror neurons in auditory perception of obstruents are more strongly activated than other phoneme classes. There is initial evidence for this line of thinking in neuroscience. Mesgarani, Cheung, Johnson, and Chang (2014) utilized an exceptional methodology called electrocorticography, which places electrodes directly on the brain while a patient undergoes evaluation for epilepsy. This type of technique uniquely offers both spatial and temporal precision that functional magnetic resonance imaging and EEGs cannot provide. This particular study showed that the auditory perception of each phoneme caused a spike at a specific location on the superior temporal gyrus, and the location of the spikes grouped by manner of articulation. Following the results from Mesgarani et al., it seems plausible that the human brain processes obstruents differently than vowels and sonorants, which could influence the outcome of an L2 training experiment that targets sounds of differing manners of articulation.

A second way that we can explain the large effect sizes of training obstruents may be the type of change that the studies targeted. Any perception training study focusing on L2 segments can be classified as having one of three foci: new category, feature adjustment, or new phonotactic constraint. An example of training a new category is a Greek speaker, who only has one high front vowel /i/, needing to create a new vowel category for a French /y/. A feature adjustment can be exemplified by an English speaker learning new VOT standards for Spanish plosives. An example of the third type is a Korean speaker learning to allow fricatives in the coda position as Arabic does; in Korean, fricatives are always followed by a vowel and can never appear as a coda. Our sample of studies was too small to support this type of breakdown in the analysis, but this question is crucial for understanding the learning of L2 speech sounds. One reason that obstruents trained better may have been that four out of five of the training programs focused on feature or phonotactic learning, and the fifth one focused on phonotactic and categorical change. None focused on new category learning alone. Future L2 training research that investigates this three-way distinction would be valuable.

Recommendations for future research

As we conducted this meta-analysis, certain aspects of experimental design became apparent from the bird's-eye view, and we consolidated these concerns into five recommendations for L2 phonology researchers, as follows.

First, *adopt standard data reporting practices.* As the field of SLA continues to develop, researchers need to hold each other accountable for adopting good, basic reporting standards. We encourage all authors to always include means and standard deviations for all tests. Twelve studies could not be included in the meta-analysis because we could not obtain simple means and standard deviations for pre- and posttests. In addition, Larson-Hall and Plonsky (2015) and Plonsky (2014) call for SLA researchers to report reliability measures. In the case of production data, calculating the reliability of human raters is important. Only 30% of the studies that we synthesized reported interrater agreement.

Second, *increase sample sizes.* The 30 studies that we synthesized had 51 unique experimental groups, and the average sample size was 11.61 participants. The smallest reported *n* was 2 (i.e., Hardison, 2003), and there were only 3 experimental groups that had an *n* larger than 20 (i.e., Anderson, 2011; Iverson et al., 2012; Underbakke, 1993). One struggle particular to L2 phonetic training is the great lengths researchers must go to in order to recruit and retain participants. However, if researchers wish to make meaningful contributions to understanding SLA and to be able to confidently generalize the findings, the domain must make an effort to substantially increase sample sizes.

Third, *increase the range of languages and target phonemes studied.* On a similar note, in order to generalize the findings of individual study results, researchers in this domain must make a concerted effort to increase the variety of languages they study and the target phonemes they choose. That one third of the synthesized studies focused on the /ɹ/–/l/ distinction shows that this phoneme contrast is popular. However, in order to gain a bigger picture of how humans learn difficult L2 sounds, it is imperative that the research community increases the spread of

https://doi.org/10.1017/S0142716417000418 Published online by Cambridge University Press

their research. The meta-analysis of L2 pronunciation reported that 97% of the studies they analyzed used English as a first or target language (Lee et al., 2015). In this report, English was the first or target language of 29 out of the 30 studies we synthesized.

Fourth, *be cautious when creating tests and tests of generalization.* Most study designs included tests of generalization to see if improvements achieved during training transferred to different testing conditions. Features that can be altered for generalization purposes include changes in task, speaker, phonetic context (both preceding/following sounds, and syllable, word, or sentential context), audiovisual presentation, or stimuli quality (e.g., synthetic or natural voice). Generalization tests are a valuable addition to address the ecological validity of any training. However, some experiments deliberately or unknowingly used a test of generalization as the primary dependent measure. Designs that use a primary test that includes generalization of some sort are less likely to find gains compared to a test that replicates all aspects of the training; furthermore, these two types of test designs will not produce gain scores that are comparable across studies. For example, Soler-Urzua (2011) utilized two perception measures that altered the task, speaker, and words from the training, and the effect size changed dramatically from –0.10 to 0.79. In order to make results more comparable across studies, we recommend that training studies use both a primary test that does not change any feature from the training and secondary tests that generalize features. In addition, only 7 out of the 30 synthesized studies administered a delayed posttest. This type of data has the potential to inform researchers and learners about the long-term effects of phonetic training programs.

Fifth and finally, *when testing a theoretical claim, be sure to strictly isolate the modalities and control all extraexperimental exposure.* Many of the studies we reviewed for inclusion in the meta-analysis stated an interest in the connection between the perception and production modalities. However, the training programs utilized procedures that confounded the two. If participants hear examples of the target phonemes and are prompted to repeat the stimulus, they are actually training both perception and production. In order to truly understand the connection between these two modalities, they need to be strictly isolated in the training: isolated perception training offers participants no opportunity to produce the target sounds, and isolated production training offers participants no opportunity to hear the target sounds. The latter type of training is not particularly intuitive and might necessitate creative thinking. Another concern is access to the target phonemes outside of the training session. Researchers should consider one-session trainings if they truly want to control for any engagement with the target phonemes outside of the experiment. If the domain is careful about controlled methodological design, there will be an abundance of varied perception-only and production-only training studies that we can analyze in another 25 years.

Conclusion

Ultimately, the present meta-analysis was able to show that perception-only training can lead to production gains. This finding is encouraging for L2 instructors and learners. It is also a valuable contribution for the SLM and for the development

https://doi.org/10.1017/S0142716417000418 Published online by Cambridge University Press

of theoretical insights into the relationship of perception and production in speech processing more generally. Within months of the conclusion of our exhaustive literature search for this meta-analysis, two more studies, Inceoglu (2014) and Rato (2013), were published that would have passed the inclusion and exclusion criteria. Both of these studies targeted L2 vowels and the effect sizes of production scores after perception training was consistent with the findings of this meta-analysis.[9] In time, more published work in this area will build a comprehensive picture of the perception–production connection and how training in one modality can influence acquisition and development of L2 sounds in both modalities.

ACKNOWLEDGMENTS

We are grateful to the reviewers who provided invaluable feedback on an earlier version of this article. We also thank Lourdes Ortega and John Norris for their guidance at the inception of the project and in later drafts of the manuscript.

NOTES

1. According to Google Scholar, Bradlow et al. (1997) had been cited by 542 articles at the time of writing this manuscript. Eight out of the first 10 studies listed use Bradlow et al. (1997) alone as evidence that the production of L2 sounds can improve after perception-only training. The remaining two cite Rochet (1995) as additional evidence.
2. Unfortunately, Hemphill (1999) and Huthaily (2008) were eliminated at this stage because the methodological design included a control group for comparison after training rather than a pre- and posttest design. Otherwise, these two studies would have passed all inclusion criteria.
3. Thank you to the researchers who graciously responded to our emails: Anderson, Baese-Berk, Brosseau-Lapré, Handley, Haslam, Iverson, Imaizumi, and Schmidt. Unfortunately, most authors were unable to provide the information necessary to calculate effect sizes. One author, Anderson, was able to supply the needed data, and we are very grateful for the time he spent communicating with us.
4. We decided to follow the suggestions of Morris (2008) regarding the use of pooled and mean pretest standard deviations in the PP and PPC effect size calculations, respectively. Please refer to this article for a more detailed discussion of how standard deviations in the denominator can influence the overall effect size calculations.
5. This category includes Sawallis and Townley (2009) and Stenning and Jamieson (2002) that were presented at a meeting of the Acoustical Society of America. The corresponding journal routinely publishes all presentation abstracts from the conference. It was decided that these studies would be included in the meta-analysis since the abstracts are published.
6. Baese-Berk (2010) used phoneme targets with varying VOTs that exist in many languages, but she did not name a particular target language.
7. If a study reported a range of hours spent during training, the median number was used to calculate the average reported here (e.g., Lengeris, 2009; Thomson, 2007). This same method was used for calculating the average number of training sessions and training days.

https://doi.org/10.1017/S0142716417000418 Published online by Cambridge University Press

8. Hazan et al. (2005) and Lambacher, Martens, Kakehi, Marasinghe, and Molholt (2005) were retained for analysis even though the participants in both studies who were tested for production gains were a subset of the population that underwent perception training. Although the perception scores reflected a larger sample, we decided that the perception effect size reasonably represented the effectiveness of the training, albeit of a larger sample than the production data.
9. Inceoglu (2014) trained two groups of participants, and the effect of perception training on production was $d = 0.30$ and 0.16. Rato (2013) trained one group, and the effect was $d = 0.38$.

REFERENCES

References marked with an asterisk indicate the studies passed all inclusion and exclusion criteria (with the exception of sufficient data reporting to calculate effect sizes) and were included in the synthesis.

Akita, M. O. (2007). Global foreign accent and the effectiveness of a prosody-oriented approach to EFL classrooms. In H. Caunt-Nulton, S. Kulatilake, & I. Woo (Eds.), *Proceedings of the annual Boston University Conference on Language Development* (pp. 46–57). Somerville, MA: Cascadilla Press.

Aliaga-Garcia, C., & Mora, J. C. (2007). Assessing the effects of phonetic training on L2 sound perception and production. In A. S. Rauber, M. A. Watkins, & B. O. A. Baptista (Eds.), *New Sounds 2007: Proceedings of the 5th International Symposium on the Acquisition of Second Language Speech* (pp. 10–27). Florianópolis, Brazil: Federal University of Santa Catarina.

Aliaga-Garcia, C., & Mora, J. C. (2009). Assessing the effects of phonetic training on L2 sound perception and production. In M. A. Watkins, A. S. Rauber, & B. O. Baptista (Eds.), *Recent research in second language phonetics/phonology: Perception and production* (pp. 2–31). Newcastle upon Tyne: Cambridge Scholars.

Altvater-Mackensen, N., & Fikkert, P. (2010). The acquisition of the stop-fricative contrast in perception and production. *Lingua, 120*, 1898–1909.

*Anderson, G. (2011). *Non-linear dynamics of adult non-native phoneme acquisition perception and production* (Unpublished doctoral dissertation).

*Baese-Berk, M. (2010). *An examination of the relationship between speech perception and production* (Unpublished doctoral dissertation).

Bettoni-Techio, M. (2008). *Perceptual training and word-initial /s/-clusters in Brazilian Portuguese English interphonology* (Unpublished doctoral dissertation).

Birdsong, D. (Ed.). (1999). *Second language acquisition and the critical period hypothesis*. Mahwah, NJ: Erlbaum.

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 and Psychophysics, 61*, 977–985.

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

*Brosseau-Lapré, F., Rvachew, S., Clayards, M., & Dickson, D. (2013). Stimulus variability and perceptual learning of nonnative vowel categories. *Applied Psycholinguistics*, *34*, 419–441.

Byun, T. M. (2012). Bidirectional perception-production relations in phonological development: Evidence from positional neutralization. *Clinical Linguistics and Phonetics, 26*, 397–413.

https://doi.org/10.1017/S0142716417000418 Published online by Cambridge University Press

Byun, T. (2015). Perceptual discrimination across contexts and contrasts in preschool-aged children. *Lingua, 160*, 38–53.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., . . . David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science, 276*, 593–596.

Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G., McGuire, P., Suckling, J., . . . David, A. S. (2001). Cortical substrates for the perception of face actions: An fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Cognitive Brain Research, 12*, 233–243.

Cenoz, J., & Garcia Lecumberri, L. (1999). The effect of training on the discrimination of English vowels. *International Review of Applied Linguistics, 37*, 261–275.

Chen, J. L., Penhune, V. B., & Zatorre, R. J. (2008). Listening to musical rhythms recruits motor regions of the brain. *Cerebral Cortex, 18*, 2844–2854.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

D'Ausilio, A., Bufalari, I., Salmas, P., & Fadiga, L. (2012). The role of the motor system in discriminating normal and degraded speech sounds. *Cortex*, *48*, 882–887.

DeKeyser, R. (2015). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 94–112). London: Routledge.

de Leeuw, E., Mennen, I., & Scobbie, J. (2013). Dynamic systems, maturational constraints, and phonetic attrition. *International Journal of Bilingualism, 17*, 683–700.

Dziubalska-Kołaczyk, K., Wrembel, M., & Kul, M. (Eds.). (2010). *New Sounds 2010: Proceedings of the sixth international symposium on the acquisition of second language speech*. Poznań, Poland: Uniwersytet im Adama Mickiewicza.

Flege, J. E. (1988a). The production and perception of speech sounds in a foreign language. In H. Winitz (Ed.), *Human communication and its disorders: A review* (pp. 224–401). Norwood, NJ: Ablex.

Flege, J. E. (1988b). Using visual information to train foreign-language vowel production. *Language Learning, 38*, 365–407.

Flege, J. (1989). Chinese subjects' perception of the word-final English /t/-/d/ contrast: Performance before and after training. *Journal of the Acoustical Society of America, 86*, 1684–1697.

Flege, J. E. (1992). Speech learning in a second language. In C. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, and implications* (pp. 565–604). Timonium, MD: York Press.

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

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

Flege, J. E. (2002). Interactions between the native and second-language phonetic systems. In P. Burmeister, T. Piske, & A. Rohde (Eds.), *An integrated view of language development: Papers in honor of Henning Wode* (pp. 217–244). Trier, Germany: Wissenschaftlicher Verlag.

Flege, J. E. (2003). Assessing constraints on second-language segmental production and perception. In A. Meyer & N. Schiller (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities* (pp. 319–355). Berlin: Mouton de Gruyter.

García-Pérez, G. M. (2003). *Training Spanish speakers in the perception and production of English vowels* (Unpublished doctoral dissertation).

García-Pérez, G. M. (2006). Production: The starting link. *Revista Virtual de Estudios da Linguagem, 4*, 1–12.

https://doi.org/10.1017/S0142716417000418 Published online by Cambridge University Press

Gervain, J., & Mehler, J. (2010). Speech perception and language acquisition in the first year of life. *Annual Review of Psychology, 61*, 191–218.

Golfeto, R. M., & de Souza, D. G. (2015). Sentence production after listener and echoic training by prelingual deaf children with cochlear implants. *Journal of Applied Behavior Analysis, 48*, 363–375.

*Gómez Lacabex, E., & García Lecumberri, M. (2010). Investigating training effects in the production of English weak forms by Spanish learners. In K. Dziubalska-Kołaczyk, M. Wrembel, & M. Kul (Eds.), *New Sounds 2010: Proceedings of the sixth international symposium on the acquisition of second language speech* (pp. 137–143). Poznań, Poland: Uniwersytet im Adama Mickiewicza.

Gómez Lacabex, E., García Lecumberri, M. L., & Cooke, M. (2009). Training and generalization effects of English vowel reduction for Spanish listeners. In M. A. Watkins, A. S. Rauber, & B. O. Baptista (Eds.), *Recent research in second language phonetics/phonology: Perception and production* (pp. 32–42). Cambridge: Cambridge Scholars.

Granena, G., & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research, 29*, 311–343.

*Hamada, M., & Tsushima, T. (2001). Relation between improvements of perception and production ability during a university English speech training course. In J. White (Ed.), *Proceedings of the Fourth Conference on Foreign Language Education and Technology* (pp. 579–582). Kobe: Japan Association for Language Education and Technology.

*Han, J.-I. (2002). The effect of formal instruction on acquisition of the English /r/ and /l/ by Korean speakers. *Language Research, 38*, 691–711.

*Handley, Z., Sharples, M., & Moore, D. (2009). *Training novel and phonemic contrasts: A comparison of identification and oddity discrimination training*. Paper presented at the 2009 ISCA Workshop on Speech and Language Technology in Education, Warwickshire, September 3–5. Retrieved from http://www.eee.bham.ac.uk/SLaTE2009/papers/SLaTE2009-02-v2.pdf

Hanlon, E. H. (2005). *The role of self-judgment and other-perception in English pronunciation attainment by adult speakers of Spanish* (Unpublished doctoral dissertation).

Hansen Edwards, J. G., & Zampini, M. L. (Eds.). (2008). *Phonology and second language acquisition*. Philadelphia, PA: John Benjamins.

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

*Haslam, M. (2011). *The effect of perceptual training including required lexical access and meaningful linguistic context on second language phonology* (Unpublished doctoral dissertation).

Hattori, K., & Iverson, P. (2010). *Examination of the relationship between L2 perception and production: An investigation of English /r/-/l/ perception and production by adult Japanese speakers*. Paper presented at the Interspeech Workshop on Second Language Studies: Acquisition, Learning, Education and Technology, Waseda University.

*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*, 360–378.

Hemphill, R. (1999). Experimental evidence for separate perception and production processing. *Papers from the Regional Meetings, Chicago Linguistic Society, 35*, 167–178.

Herd, W. (2011). *The perceptual and production training of /d, tap, r/ in L2 Spanish: Behavioral, psycholinguistic, and neurolinguistic evidence* (Unpublished doctoral dissertation).

*Herd, W., Jongman, A., & Sereno, J. (2013). Perceptual and production training of intervocalic /d, ɾ, r/ in American English learners of Spanish. *Journal of the Acoustical Society of America, 133*, 4274–4255.

*Huensch, A. (2013). *The perception and production of palatal codas by Korean L2 learners of English* (Unpublished doctoral dissertation).

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.

Huthaily, K. Y. (2008). *Second language instruction with phonological knowledge: Teaching Arabic to speakers of English* (Unpublished doctoral dissertation).

*Imaizumi, S., Itoh, H., Tamekawa, Y., Deguchi, T., & Mori, K. (1998). Plasticity of non-native phonetic perception and production: A training study. In R. H. Mannell & J. Robert-Ribes (Eds.), *Proceedings of the international conference on spoken language processing* (pp. 1887–1890). Sydney: Australian Speech Science and Technology Association.

Inceoglu, S. (2014). *Effect of multimodal training on the perception and production of French nasal vowels by American English learners of French* (Unpublished doctoral dissertation).

Iverson, P., & Evans, B. G. (2009). Learning English vowels with different first-language vowel systems: II. Auditory training for native Spanish and German speakers. *Journal of the Acoustical Society of America, 126*, 866–877.

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

*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*, 145–160.

James, A. R. (Ed.). (1994). Second language phonology *anno* 1994 [Special issue]. *Second Language Research, 10*(3).

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. *Journal of the Acoustical Society of America, 138*, 817–832.

Kissling, E. (2012). *The effects of phonetics instruction on adult learners' perception and production of L2 sounds* (Unpublished doctoral dissertation).

Kondaurova, M. V., & Francis, A. L. (2010). The role of selective attention in the acquisition of English tense and lax vowels by native Spanish listeners: Comparison of three training methods. *Journal of Phonetics, 38*, 569–587.

Lado, R. (1989). Acquisition versus learning in reading pronunciation by adult EFL students. In J. Alatis (Ed.), *Georgetown University round table on language and linguistics 1989* (pp. 59–68). Washington, DC: Georgetown University Press.

Lambacher, S., Martens, W., & Kakehi, K. (2002). The influence of identification training on identification and production of the American English mid and low vowels by native speakers of Japanese. In J. H. L. Hansen & B. Pellom (Eds.), *Seventh International Conference on Spoken Language Processing* (pp. 245–248). Denver, CO: International Speech Communication Association.

*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*, 227–247.

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning, 65*, 127–159.

Leather, J. (Ed.). (1999a). Second-language speech research [Special issue]. *Language Learning*, *49*(1).

Leather, J. (1999b). Second-language speech research: An introduction. *Language Learning, 49*, 1–56.

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

Lengeris, A. (2008). The effectiveness of auditory phonetic training on Greek native speakers' perception and production of Southern British English vowels. *Experimental Linguistics*, *133*.

*Lengeris, A. (2009). *Individual differences in second-language vowel learning* (Unpublished doctoral dissertation).

Lenneberg, E. H. (1967). *Biological foundations of language*. Oxford: Wiley.

Lipsey, M. W., & Wilson, D. (2001). *Practical meta-analysis*. Los Angeles: Sage.

https://doi.org/10.1017/S0142716417000418 Published online by Cambridge University Press

Lively, S. E., Pisoni, D. B., & Logan, J. S. (1991). Training Japanese listeners to identify English /r/ and /l/. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, production and linguistic structure* (pp. 175–196). Amsterdam: IOS Press.

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

Long, M. (1990). Maturational constraints on language development. *Studies in Second Language Acquisition*, *12*, 251–285.

Lopez-Soto, T., & Kewley-Port, D. (2009). Relation of perception training to production of codas in English as a second language. *Journal of the Acoustical Society of America*, *125*, 2756.

Lopez-Soto, T., & Kewley-Port, D. (2010). Influence of perceptual training of syllable codas for English consonants on sentences. *Journal of the Acoustical Society of America, 127*, 1954.

Major, R. C. (1998). Interlanguage phonetics and phonology: An introduction [Special issue]. *Studies in Second Language Acquisition, 20*, 131–137.

McCandliss, B. D., Fiez, J. A., Protopapas, A., Conway, M., & McClelland, J. L. (2002). Success and failure in teaching the [r]-[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective, and Behavioral Neuroscience*, *2*, 89–108.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science, 343*, 1006–1010.

Meyer, A. S., Huettig, F., & Levelt, W. J. M. (2016). Same, different, or closely related: What is the relationship between language production and comprehension? *Journal of Memory and Language, 89*, 1–7.

Mora, J. C., & Nadeu, M. (2012). L2 effects on the perception and production of a native vowel contrast in early bilinguals. *International Journal of Bilingualism*, *16*, 484–500.

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods, 11*, 354–386.

*Motohashi, M. (2007). *Acquisition of geminate consonants in Japanese by American English speakers* (Unpublished doctoral dissertation).

Motohashi-Saigo, M., & Hardison, D. M. (2009). Acquisition of L2 Japanese geminates: Training with waveform displays. *Language Learning and Technology, 13*, 29–47.

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

*Nobre-Oliveira, D. (2007). Effects of perceptual training on the learning of English vowels in non-native settings. In A. S. Rauber, M. A. Watkins, & B. O. Baptista (Eds.), *New Sounds 2007: Proceedings of the fifth international symposium on the acquisition of second language speech* (pp. 382–389). Florianópolis, Brazil: Federal University of Santa Catarina.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*, 417–528.

Patkowski, M. (1994). The critical age hypothesis and interlanguage phonology. In M. Yavas (Ed.), *First and second language phonology* (pp. 205–221). San Diego, CA: Singular.

Piske, T., MacKay, I. R., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics, 29*, 191–215.

Plonsky, L. (2014). Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform. *Modern Language Journal, 98*, 450–470.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912.

Pruitt, J. S. (1995). Perceptual training on Hindi dental and retroflex consonants by native English and Japanese speakers. *Journal of the Acoustical Society of America, 97*, 3417–3418.

https://doi.org/10.1017/S0142716417000418 Published online by Cambridge University Press

Pruitt, J. S., Jenkins, J. J., & Strange, W. (2006). Training the perception of Hindi dental and retroflex stops by native speakers of American English and Japanese. *Journal of the Acoustical Society of America, 119*, 1684–1696.

Rato, A. A. (2013). *Cross-language perception and production of English vowels by Portuguese learners: The effects of perceptual training* (Unpublished doctoral dissertation).

Rauber, A. S., Watkins, M. A., & Baptista, B. O. (Eds.). (2007). *New Sounds 2007: Proceedings of the 5th-International Symposium on the Acquisition of Second Language Speech*. Florianópolis, Brazil: Federal University of Santa Catarina.

*Reis, M., & Nobre-Oliveira, D. (2007). Effects of perceptual training on the identification and production of English voiceless plosives aspiration by Brazilian EFL learners. In A. S. Rauber, M. A. Watkins, & B. O. A. Baptista (Eds.), *New Sounds 2007: Proceedings of the Fifth International Symposium on the Acquisition of Second Language Speech* (pp. 398–407). Florianópolis, Brazil: Federal University of Santa Catarina.

*Rochet, B. L. (1995). Perception and production of L2 speech sounds by adults. In W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research* (pp. 379–410). Timonium, MD: York Press.

Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin, 86*, 638–641.

Rvachew, S., Nowak, M., & Cloutier, G. (2004). Effect of phonemic perception training on the speech production and phonological awareness skills of children with expressive phonological delay. *American Journal of Speech-Language Pathology, 13*, 250–263.

Saito, K. (2011). Examining the role of explicit phonetic instruction in native-like and comprehensible pronunciation development: An instructed SLA approach to L2 phonology. *Language Awareness*, *20*, 45–59.

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, and Psychophysics, 71*, 1207–1218.

*Sawallis, T. R., & Townley, M. W. (2009). Adapting second language phonemic perception training to common instructional situations: Pitfalls and progress. *Journal of the Acoustical Society of America, 125*, 2777.

Scovel, T. (2000). A critical review of the critical period research. *Annual Review of Applied Linguistics, 20*, 213–223.

Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: Motor cortical activation during speech perception. *NeuroImage*, *25*, 76–89.

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

Song, J. J., Lee, H. J., Kang, H., Lee, D. S., Chang, S. O., & Oh, S. H. (2015). Effects of congruent and incongruent visual cues on speech perception and brain activity in cochlear implant users. *Brain Structure and Function, 220*, 1109–1125.

*Stenning, K., & Jamieson, D. G. (2002). Improving English vowel perception and production by Spanish-speaking adults. *Journal of the Acoustical Society of America*, *112*, 2251.

Sterne, J. A., & Harbord, R. M. (2004). Funnel plots in meta-analysis. *Stata Journal*, *4*, 127–141.

Stölten, K., Abrahamsson, N., & Hyltenstam, K. (2015). Effects of age and speaking rate on voice onset time. *Studies in Second Language Acquisition*, *37*, 71–100.

Strange, W. (1995). *Speech perception and linguistic experience: Issues in cross-language research.* Timonium, MD: York Press.

*Thomson, R. I. (2007). *Modeling L1/L2 interactions in the perception and production of English vowels by Mandarin L1 speakers: A training study* (Unpublished doctoral dissertation).

Thomson, R. I. (2012). Improving L2 listeners’ perception of English vowels: A computer-mediated approach. *Language Learning*, *62*, 1231–1258.

*Todaka, Y. (2008). Receptive and productive skills of English /l/ and /r/ by Japanese college students in relation to their motivation. In A. Botinis (Ed.), *International Speech Communication*

Association proceedings of ISCA tutorial and research workshop on experimental linguistics (pp. 213–216). Athens: University of Athens.

Tohkura, Y. I., Vatikiotis-Bateson, E., & Sagisaka, Y. (1992). *Speech perception, production and linguistic structure*. Tokyo: Ohmsha.

Tseng, S., Kuei, K., & Tsou, P. (2011). Acoustic characteristics of vowels and plosives affricates of Mandarin-speaking hearing-impaired children. *Clinical Linguistics and Phonetics, 25*, 784–803.

Tuller, B. (2004). Categorization and learning in speech perception as dynamical processes. In M. A. Riley & G. C. Van Orden (Eds.), *Tutorials in contemporary nonlinear methods for the behavioral sciences* (pp. 353–400). Arlington, VA: National Science Foundation.

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

*Wang, X. (2002). *Training Mandarin and Cantonese speakers to identify English vowel contrasts: Long-term retention and effects on production* (Unpublished doctoral dissertation).

Weiss, B. (1992). Perception and production in accent training. *Revue de Phonetique Appliquee, 102*, 69–82.

Werker, J. F., & Hensch, T. K. (2015). Critical periods in speech perception: New directions. *Psychology*, *66*, 173.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, *7*, 701–702.

Wong, P., Schwartz, R. G., & Jenkins, J. J. (2005). Perception and production of lexical tones by 3-year-old, Mandarin-speaking children. *Journal of Speech, Language, and Hearing Research, 48*, 1065–1079.

*Yeon, S-H. (2004). *Teaching English word final alveolopalatals to native speakers of Korean* (Unpublished doctoral dissertation).

Yeon, S.-H. (2008). Training English word-final palatals to Korean speakers of English. *Applied Language Learning, 18*, 51–61.

Zampini, M. L. (1998). The relationship between the production and perception of L2 Spanish stops. *Texas Papers in Foreign Language Education, 3*, 85–100.

https://doi.org/10.1017/S0142716417000418 Published online by Cambridge University Press