Applied Psycholinguistics **38** (2017), 371–393
doi:10.1017/S0142716416000254

Can corrective feedback on second language speech perception errors affect production accuracy?

ANDREW H. LEE and ROY LYSTER
McGill University

Received: August 12, 2015 Accepted for publication: May 16, 2016

ADDRESS FOR CORRESPONDENCE
Andrew H. Lee, Department of Integrated Studies in Education, McGill University, 3700 McTavish Street, Montreal, QC H3A 1Y2, Canada. E-mail: andrew.lee@mcgill.ca

ABSTRACT
This study investigated whether different types of corrective feedback (CF) in second language speech perception training have differential effects on second language speech production. One hundred Korean learners of English were assigned to five different groups and participated in eight computer-assisted perception training sessions focusing on English vowels. While no CF was provided to the control group, participants in the four treatment groups received one of three types of auditory CF or a visual type of CF. A pretest, an immediate posttest, and a delayed posttest each consisted of a production measurement at a controlled-speech level. Results revealed that the extent to which the participants' production accuracy benefited from the perception training depended on CF type. In addition, by adopting the perception accuracy data by Lee and Lyster (2016b), the current study found that improvement in perception accuracy was a significant predictor of improvement in production accuracy.

It is well known that second language (L2) learners have difficulty perceiving and producing certain nonnative phonemes. Because of these acquisitional difficulties, a number of training studies have investigated whether L2 learners can overcome the difficulties, the results of which have revealed positive training effects in laboratory and classroom settings (Lambacher, Martens, Kakehi, Marasinghe, & Molholt, 2005; Lee & Lyster, 2016a; Saito, 2013). Of particular interest is the association between speech production and perception whereby L2 learners' perception accuracy serves as a predictor of their production accuracy. Several empirical studies (e.g., Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Hardison, 2003; Thomson, 2011; Wang, Jongman, & Sereno, 2003) have found that L2 learners' production accuracy increases as a result of perception training without any explicit production training. Considering the importance of L2 speech perception and its training in L2 phonological development, the previous literature leads us to the following pedagogical question: what are the most effective training techniques in L2 speech perception training?

Bradlow (2008) and Hardison (2012) argued that successful training involves stimulus variability (e.g., multiple talkers and multiple exemplars) and corrective feedback (CF) during training. Similarly, Wang and Munro (2004) also emphasized the roles of CF in their L2 speech perception training study. However, CF per se as a technique has unfortunately been out of the spotlight in the field of speech perception training. Recently, Lee and Lyster (2016a) conducted an experimental study in which they employed CF as a key technique during perception training and provided empirical evidence for its effectiveness. Furthermore, while criticizing the use of simple and uniform types of CF (i.e., right or wrong) commonly employed in the field, Lee and Lyster (2016b) examined the use of various types of CF and revealed differential effects for different CF types during L2 speech perception training.

Acknowledging that the ultimate goal of L2 speech perception training is to improve L2 learners' production as well as perception performance, Hardison (2012) argued that successful perception training should transfer to other skills such as production. Given that CF has the potential to play a key role in successful perception training as demonstrated by Lee and Lyster (2016a, 2016b), it is timely to probe the extent to which CF treatment facilitates skill transfer (i.e., from receptive skills to productive skills). Accordingly, the current study aims to explore (a) whether L2 learners' production performance benefits from perception training without any explicit production training and, (b) if so, whether the benefits depend on perception training techniques, particularly different types of CF. In addition, given that the participants in the current study were the same learners as in Lee and Lyster (2016b), which investigated their changes in perception accuracy by CF type, the present study attempts to delve into the relationship between improvement in perception accuracy and improvement in production accuracy by adopting the perception accuracy data from Lee and Lyster (2016b).

By investigating whether the effects of perception training on production accuracy vary depending on CF type, the current study is expected to provoke methodological debates concerning L2 speech perception training and to expand horizons in regard to the roles attributed to CF. In addition, the present study is expected to provide empirical evidence with respect to the relationship between L2 speech production and perception.

BACKGROUND

Perception-first view in L2 phonological acquisition

In the realm of L2 phonological acquisition, a relationship between L2 speech production and perception has been postulated, which in turn has been both theoretically and empirically addressed in terms of the following questions: Are the two modalities connected? If so, does one precede the other? Although not espousing a perception-first view explicitly, the perceptual assimilation model (PAM, Best, 1995; PAM-L2, Best & Tyler, 2007) posits that speakers employ articulatory gestures as the basis of speech perception. Furthermore, the speech learning model (SLM; Flege, 1995; Flege, Schirru, & MacKay, 2003) puts forward that well-formed phonological representation at the perception level is a sine qua non for targetlike sensory motor skills and accurate L2 speech production. In particular, SLM hypothesizes that nonnative phonemes are categorized as either new sounds or similar sounds based on a L2 learner's first language (L1) phonemic inventory. The model predicts that L2 learners, therefore, need more time and intervention to acquire similar sounds compared to new sounds.

The perception-first view in L2 phonological acquisition has been supported by many empirical studies such as those by Bradlow et al. (1997), Hardison (2003), Thomson (2011), and Wang et al. (2003), which showed that the production accuracy of L2 learners benefited from perception training without any production-related training. However, some studies (de Jong, Hao, & Park, 2009; Goto, 1971; Sheldon & Strange, 1982) challenged the relationship between L2 speech production and perception. In particular, Baker, Trofimovich, Flege, Mack, and Halter (2008) argued that inaccurate perceptual performance is more likely a matter for L2 beginners. In a similar vein, according to Kissling (2014) and Strange and Shafer (2008), the discrepant findings might result from methodological issues or from other variables that influence L2 speech production (e.g., time spent using the L2 outside the classroom, age, and attitude). In this respect, it is less contentious in the field of study that there is a link between L2 speech production and perception at certain levels and that perception precedes production (for details, see Colantoni & Steele, 2008). Neurological studies have also underpinned this claim (Pulvermüller & Shumann, 1994; Watkins & Paus, 2004).

It is arguable that L2 learners difficulty perceiving certain nonnative phonemes is not due to basic auditory capabilities. Rather, it may be the case that certain phonetic differences (i.e., acoustic differences) result in phonological differences (i.e., differences in meaning) in one language, whereas the same phonetic differences do not necessarily result in phonological differences in the other language. According to Strange's (2006, 2007) automatic selective perception model, infants are pushed to tune into acoustic cues, which are crucial for processing their L1 while ignoring other irrelevant cues. According to this account, online L1 speech perception by adults is processed via highly overlearned selective perceptual routines (SPRs).

When L2 learners perceive L2 phonemes, the SPRs predispose them to select and integrate acoustic cues that are crucial in processing their L1, but not with respect to their L2. Consequently, L2 learners fail to perceive L2 phonemes correctly. In spite of such perceptual difficulties, it is possible for L2 learners to create and develop SPRs that are more targetlike over their life span. For instance, Flege's SLM (Flege, 1995; Flege et al., 2003) and related empirical research (e.g., Flege, 2002) suggest that the ability to modify L2 speech perception patterns is maintained well into adulthood. As Bradlow (2008) explained, however, the "retuning" procedure to develop targetlike perception patterns is not something that occurs in an incidental manner. Instead, it appears to require a great deal of language exposure and explicit training.

As for explicit training, previous research has investigated the effects of several training techniques, mostly in laboratory settings and particularly via computer-assisted perception training. For example, high-variability phonetic training methods, such as multiple talkers (Lively, Logan, & Pisoni, 1993), auditory–visual stimuli (Hardison, 2003), and hyperarticulated/exaggerated stimuli (Iverson, Hazan, &

Bannister, 2005; Uther, Knoll, & Burnham, 2007) have all been tested in laboratory settings. A fading stimulus presentation scheme (e.g., Pruitt, 1995) and various training tasks such as oddity discrimination, categorical discrimination, and identification tasks (Strange & Shafer, 2008) were also adopted in laboratory-based training studies. Outside of a typical laboratory setting, Lee and Lyster (2016a) conducted perception training in a simulated classroom context, employing several pedagogical techniques such as awareness tasks and relevant L2 instruction. A close look at the literature in all these cases helped us to find one common training technique, namely, CF intervention in response to L2 learners' errors, leading to the question: what are the roles of CF in L2 learning?

Roles of CF in L2 learning

The effectiveness of CF in L2 learning has been demonstrated in many studies (for meta-analysis and narrative reviews, see Li, 2010; Lyster & Saito, 2010; Lyster, Saito, & Sato, 2013). The learning mechanisms of CF are explained by several L2 acquisition theories such as skill acquisition theory (DeKeyser, 1998, 2001; Lyster & Sato, 2013), the output hypothesis (Swain, 1985, 1995), and the interaction hypothesis (Long, 1996). With respect to different types of CF, for example, there are two main groups in the CF literature: one provides L2 learners with the target forms (e.g., recasts) and the other withholds the target forms and pushes L2 learners to produce modified output (e.g., prompts). Recasts are considered effective because they provide L2 learners with opportunities to notice the gap between interlanguage forms and target forms, whereas prompts are considered effective because they provide L2 learners with signals to retrieve target forms on their own and thus to engage in practice opportunities that lead to a restructuring of their L2 interlanguage system. The absence of any type of CF is thought to prevent L2 learners from attaining more targetlike accuracy, as their interlanguage representations become automatized routines.

Notwithstanding the richness of the CF literature, a disproportionate number of studies have investigated the effects of CF on productive skills (i.e., spoken and written outcomes), exclusively with regard to morphosyntatic errors. As Lyster et al. (2013) concluded, the CF field would benefit from investigating the effects of different CF types on various linguistic targets such as L2 phonological and pragmatic features. Ellis, Loewen, and Erlam (2006) and McDonough (2007) also suggested that the field aim to explore differential learning outcomes resulting from different types of CF. In this sense, a recent study by Lee and Lyster (2016b) investigated the effects of CF on L2 speech perception training (i.e., a domain of receptive rather than productive skills) and compared four different CF types motivated by the existing CF literature (Lyster et al., 2013). In their study, 100 Korean learners of English were assigned to five different groups (i.e., four different CF types plus a no-CF control) and participated in eight computer-assisted perception training sessions. During the training sessions, whereas the control group received no CF, participants in each treatment group received one of three auditory CF types (a rejection followed by the target form; a rejection followed by the learner's nontarget form; or a rejection followed by both the target and nontarget forms), and a fourth group received a *wrong* visual type of CF. Overall, the CF treatment groups outperformed the control group at the immediate and delayed posttests. Within the CF groups, the auditory CF groups performed better than the *wrong* visual CF group, which failed to outperform the control group regarding trained words recorded by unfamiliar voices. Among the auditory CF groups, the group given auditory CF contrasting both target and nontarget forms showed the highest performance.

RESEARCH QUESTIONS

Successful perception training should enable L2 learners to improve not only their perception but also their production accuracy (e.g., Hardison, 2012). In the same vein, while acknowledging the effectiveness of CF in perception training, Lee and Lyster (2016a) speculated that "CF may affect L2 learners' perception first and may then trigger successful production of L2 sounds" (p. 36). In order to test this speculation and investigate whether the benefits of L2 speech perception training implemented with various types of CF extend to L2 speech production, the current study extends Lee and Lyster's (2016b) study to investigate the following research questions:

1. As a result of computer-assisted perception training targeting the English vowel pairs /i/–/ɪ/ and /ɛ/–/æ/, to what extent do Korean learners of English improve their production accuracy regarding the target pairs in perceptually trained words and in perceptually untrained words?
2. To what extent do the perception training effects on production accuracy differ according to three different types of auditory CF (i.e., rejection plus target form; rejection plus nontarget form; and rejection plus target and nontarget forms) and one type of visual CF (i.e., "wrong" appears on computer screen)?
3. Does improvement in perception accuracy resulting from computer-assisted perception training predict improvement in production accuracy?

Because the participants in the current study are the same as those in Lee and Lyster (2016b), the perception accuracy data from the earlier study are used here in order to answer research question 3.

METHODOLOGY

Participants

Participants in the current study consisted of 100 Korean learners of English (73 females, 27 males), with a mean age of 30.3 ($SD = 9.69$). They were the same participants who had participated in Lee and Lyster (2016b). Most of them were students learning English as a foreign language in private or university language institutions, while residing in the Montreal area. Their average length of residence in English-speaking countries was 21.8 months ($SD = 18.1$), 19.1 months ($SD = 19.5$), 20.1 months ($SD = 18.8$), 18.9 months ($SD = 19.7$), and 22.1 months ($SD = 18.1$) in each of the five different groups. A majority of them had lived abroad after the age of 18. In addition, they self-reported that they were intermediate or advanced learners based on their length of learning experience in formal instructional settings ($M = 9.8$ years, $SD = 8.62$).

A total of 64 native speakers of English (50 females, 14 males) also participated in the study, serving as L1 speakers for stimuli ($n = 4$), L1 baseline participants ($n = 20$), and L1 raters for speech tokens recorded by the L1 baseline and L2 participants ($n = 40$). All the English native participants were L1 English speakers residing in the Montreal area with a mean age of 22.9 ($SD = 5.23$). They were North American English speakers coming from the United States and English-speaking provinces in Canada and were attending universities in Montreal, Canada. Those who were recruited as the L1 raters did not have any background in linguistics or applied linguistics and were thus considered inexperienced raters in the current study.

Procedure

The 100 Korean learners of English took a pretest, and the 20 L1 baseline participants completed a baseline test at the research laboratory. After the pretest, the Korean participants were randomly assigned to one of four CF treatment groups (three auditory CF groups, one visual CF group) or the control group (20 participants per group) in which they participated in eight computer-assisted perception training sessions scheduled over a 2-week period. Those who were in the four CF treatment groups received a specific type of CF when they made perceptual errors during the perception training sessions:

- the target group heard a rejection followed by the target form,
- the nontarget group heard a rejection followed by the nontarget form,
- the combination group heard a rejection followed by the target and nontarget forms, and
- the wrong group saw “wrong” on computer screen.

Those who were in the control group participated in the same perception training; however, they did not receive any CF when they made any perception errors. An immediate posttest was conducted on the last day of the training sessions, and 2 weeks later a delayed posttest was administered. All the tests were designed to measure the participants’ production accuracy at a controlled-speech level. Specifically, to measure their production accuracy, speech tokens were judged by the 40 L1 raters to assess the extent to which the perception training including different types of CF improved the L2 participants’ production accuracy.

Target stimuli

Stimuli for the perception training were prepared with the aid of the four native speakers of English. Vowels are known to play an important role in intelligibility (Bent, Bradlow, & Smith, 2007), and previous studies have shown that L2 learners experience more difficulty acquiring L2 vowels than L2 consonants (Munro & Derwing, 2008; Neri, Cucchiarini, & Strik, 2006). Nevertheless, there is a general

Table 1. *Target stimuli*

Measurement	/i/	/ɪ/	/ɛ/	/æ/
		Trained Words		
Perception	Cheap	Chip	Beg	Bag
	Feet	Fit	Bet	Bat
	Heat	Hit	Head	Had
	Meal	Mill	Hem	Ham
	Reach	Rich	Mess	Mass
	Scene	Sin	Peck	Pack
	Seek	Sick	Set	Sat
	Wheel	Will	Shell	Shall
Perception + production	*Beat	*Bit	*Bed	*Bad
	*Leave	*Live	*Dead	*Dad
	*Peel	*Pill	*Guess	*Gas
	*Sheep	*Ship	*Said	*Sad
		Untrained Words		
Perception	Cheek	Chick	Neck	Knack
	Lead	Lid	Pet	Pat
	Seal	Sill	Vet	Vat
Perception + production	*Feel	*Fill	*Leg	*Lag
	*Heal	*Hill	*Pen	*Pan
	*Peak	*Pick	*Ten	*Tan

lack of research targeting vowels in L2 speech perception training (Thomson, 2011). The current study thus targeted two pairs of English vowels (/i/–/ɪ/ and /ɛ/–/æ/). Several speech learning theories such as the PAM-L2 (Best & Tyler, 2007) and the SLM (Flege, 1995; Flege et al., 2003) predict that Korean learners of English will encounter difficulties regarding the target pairs owing to their absence in the Korean vowel inventory (Lee, 1993). For example, according to PAM-L2, each target pair might show an assimilation pattern of single-category or category-goodness difference, and Korean learners of English will thus have difficulty categorizing each vowel in each target pair. In addition, the SLM hypothesizes that the English vowel pairs /i/–/ɪ/ and /ɛ/–/æ/ are each somewhat similar to the Korean phonemes /i/ and /ɛ/ and, therefore, that Korean learners of English need more L2 experience and intervention to acquire these two contrasts properly. These difficulties have been confirmed by a number of empirical studies (e.g., Baker, Trofimovich, Mack, & Flege, 2002; Flege, Bohn, & Jang, 1997; Ingram & Park, 1997; Tsukada et al., 2005). Tsukada et al. (2005), for example, revealed that native speakers of English identified Korean adults' productions of the target pairs as various phonemes: for instance, the productions of English /i/ and /ɪ/ as /i/, /ɪ/, /e/, or /ɛ/, and the productions of English /ɛ/ and /æ/ as its counterpart, respectively.

Eighteen sets of English minimal pairs with /i/–/ɪ/ and another 18 sets of English minimal pairs including /ɛ/–/æ/ were targeted (see Table 1). These stimuli are the

Table 2. *Mean (standard deviation) formant values (Hz) for each target vowel according to gender*

	/i/		/ɪ/		/ɛ/		/æ/	
Formants	Male	Female	Male	Female	Male	Female	Male	Female
*F*0	120	253	119	256	123	248	129	241
	(16.2)	(19.8)	(18.2)	(20.1)	(16.3)	(21.3)	(18.2)	(24.2)
*F*1	295	384	379	521	502	650	604	870
	(33.5)	(41.2)	(36.2)	(35.2)	(42.3)	(42.3)	(62.1)	(65.2)
*F*2	2155	2442	1664	2164	1662	1829	1492	1897
	(98.2)	(131.2)	(121.2)	(121.7)	(78.2)	(162.2)	(85.3)	(189.2)

same as those used in Lee and Lyster (2016b). The target stimuli were monosyllabic consonant–vowel–consonant English words with various onsets and codas. They were selected from the Corpus of Contemporary American English (Davies, 2008). Except for the words underlined in Table 1, all the words were of high frequency in the corpus (>15 occurrences/million words). In addition, in order to control for any influence of word unfamiliarity, the Korean participants were asked which words they knew listed in Table 1 prior to the pretest. The first author then explained, in Korean, the meaning of the words that they did not know and modeled their pronunciation.

In Table 1, the trained words refer to the items occurring during the perception training, whereas the untrained words are those that did not occur during the perception training. The word pairs with asterisks (4 of the 12 pairs of trained words, 3 of the 6 pairs of untrained words) were employed in the production ratings to determine whether the perception training would improve the L2 learners' production of the target pairs.

The four English L1 speakers for stimuli (two males: M1 and M2; two females: F1 and F2) were asked to read each stimulus twice embedded in a carrier sentence "I said [*X*]" after which the target stimuli were extracted from the carrier sentences and digitalized at 44,100 Hz using Praat (Boersma & Weenink, 2013). One out of two productions from each speaker, considered a good exemplar, was selected by the first author and research assistants. The final stimuli were acoustically analyzed (see Table 2), which confirmed that they were valid samples compatible with previous acoustic studies (e.g., Hillenbrand, Getty, Clark, & Wheeler, 1995; Yang, 1996).

Perception training

Based on its attested effectiveness (Wang & Munro, 2004), forced-choice identification training was implemented in the current study. The training was delivered through a computer-assisted perception training program, which was specifically designed for the purposes of this research by using programmed web-based computer scripts that included JavaScript, PHP, and MySQL (Nixon, 2012). The perception training was operationalized as individual training with a computer. A minimal pair of stimuli and a relevant sound file were provided to the L2 participants. For example, a sound file conveying the word "ship" played after which two orthographic options (*sheep* and *ship*) appeared on the computer screen. The L2 participants were in turn asked to answer the question "What did s/he say?" (i.e., to select what they heard). The participants were allowed to take as much time as they needed to answer the question. In addition, because there was a *repeat* button available, they were also permitted to listen to the sound file repeatedly if they wished. Once they chose the answer, the appropriate CF treatment was provided depending on their answer and group; then, the next trial ensued. For those who were in the control group, however, the next trial was automatically provided as soon as they selected their answer.

For the CF intervention, Speakers M1 and F1 and Speakers M2 and F2 were paired. For example, if Speaker M1 provided the stimulus, Speaker F1 provided the relevant CF, and vice versa. Speakers M2 and F2 were matched in the same manner. When learners chose the wrong answer, different types of CF were provided in each treatment group. The following describes each type of CF for the four treatment groups, using the example of a learner selecting "sheep" in response to the auditory stimulus "ship":

- Target group: "*No, s/he said 'ship.'*"
- Nontarget group: "*No, not 'sheep.'*"
- Combination group: "*No, s/he said 'ship,' not 'sheep.'*"
- Wrong group: A word card with "*wrong*" written in red appears on the screen.

After the CF intervention, a pop-up message in the form of "okay?" was immediately shown on the screen with the intention of helping the L2 learners to notice their error and CF. The learners needed to click "yes" to move onto the next trial. If the learner chose the right answer (i.e., "ship" in the above example), positive oral confirmation saying "yes" was given to those in the three auditory CF groups and the word "right" appeared in blue on the screen for those in the wrong group. The next trial ensued right after the positive affirmation.

One training session comprised a total of 384 trials: the 48 trained words recorded by Speakers M1, M2, F1, and F2 with one repetition. All the trials were randomized for each participant in each training session. The L2 learners were asked to complete eight training sessions within a 2-week period, each of which took approximately 1 to 1.5 hr. Finally, it is important to note that the current study was designed to investigate the benefits of L2 speech perception training with different types of CF on L2 learners' production. Therefore, the training per se did not involve any production training; none of the tasks in the training required the L2 learners to produce the target words orally.

Production measurement

The L2 learners were audiorecorded on three occasions: in a pretest, an immediate posttest, and a delayed posttest. One single instrument was utilized to elicit the participants' productions. The learners were prompted to produce English sentences orally using the 28 target words listed in Table 1 with asterisks. Their utterances were audiorecorded in the research office in a researcher–participant dyadic setting. The first author or research assistants showed seven worksheets one by one, each of which had 4 target words, and then asked the learners to produce utterances using a carrier sentence "I said [*X*]." The learners were asked to produce the target words only once in each recording session, to equalize the number of utterances across participants. If they articulated the given words more than once, the first utterance was taken into account for the analysis. If they missed any words, the first author or research assistants induced them to produce utterances with the missing words at the end of each test.

Finally, the 20 L1 baseline participants took the same production test as the L2 learners, in order to obtain a native-speaker sample (baseline) for subsequent ratings. The purpose of having the L1 baseline participants was to ensure that the judgment results from the 40 inexperienced raters were reliable (i.e., L1 baseline participants were expected to receive consistently high ratings while variability was expected in the ratings of L2 participants).

Native-speaker judgments

In order to measure the intelligibility and comprehensibility of the target vowel pairs, the 40 inexperienced L1 raters listened individually to the speech tokens produced by the L1 baseline and L2 participants. The 28 target words produced by each participant at each test had been extracted from each participant's utterances and then digitalized at 44,100 Hz using Praat (Boersma & Weenink, 2013). As a result, a total of 8,960 speech tokens (28 words × 100 L2 participants × 3 testing sessions = 8,400, 28 words × 20 L1 baseline participants = 560) were collected. The 8,960 speech tokens were randomly divided into 8 blocks (1,120 tokens per block). Each block was judged by five different raters (8 blocks × 5 raters = 40 inexperienced L1 raters). Each block was again divided into Day 1 and Day 2 subblocks (560 tokens/day), which were judged by the same five raters on 2 separate days within a week. Each daily task took approximately 1 hr to complete.

The judgment tasks were designed using Praat (Boersma & Weenink, 2013) in order to measure intelligibility and comprehensibility of the speech tokens as suggested by Derwing and Munro (2015). The L1 raters were asked to categorize a given sound (i.e., intelligibility measurement) and then grade its goodness (i.e., comprehensibility measurement). For instance, a sound file intended as the word "ship" was played and then three options (*sheep*, *ship*, and *not sure/neither*) appeared on the computer screen. Once the raters selected one of the three options, a 5-point Likert scale appeared along with the following instruction: "Judge how good the pronunciation is between 1 (*difficulty to understand*) and 5 (*easy to understand*)." There was a *repeat* button available; the raters were allowed to change their choices until they clicked "next" to judge the next sound files.

In order to quantify the participants' production accuracy, one score was computed from the intelligibility and comprehensibility measurement tasks for each speech token employing the following coding scheme: considering the above example "ship," if a rater selected the intended word (i.e., choosing *ship*), its comprehensibility measurement (i.e., from 1 to 5) was recorded as its accuracy score. However, 0 was given regardless of its comprehensibility measurement if a rater did not choose the intended word in the intelligibility measurement (i.e., choosing *sheep* or *not sure/neither*).

Interrater reliability and data verification/preparation for statistical analyses

The Cronbach α was calculated in order to verify interrater agreement among the 40 inexperienced raters. Within the perceptually trained words, 0.82, and 0.78 were the α values for the target pairs /i/–/ɪ/ and /ɛ/–/æ/, respectively. With respect to the perceptually untrained words, values of 0.75 and 0.73 were obtained for each target pair. The reliability indexes were considered acceptable, following the benchmark value of 0.70–0.80 in L2 research studies (Larson-Hall, 2010). Thus, by averaging all raters' scores, one mean score for each target pair in each word context (i.e., the trained and untrained words) at each test was computed as the production accuracy for each participant. Using the standardized *z* score of 3.29, it was confirmed that there were no outliers in the data set. In addition, the 20 L1 baseline participants showed ceiling effects for the target pairs, /i/–/ɪ/ ($M = 4.37$, $SD = 0.75$ in the trained words; $M = 4.56$, $SD = 0.71$ in the untrained words) and /ɛ/–/æ/ ($M = 4.21$, $SD = 0.81$ in the trained words; $M = 4.29$, $SD = 0.89$ in the untrained words), thus confirming that the judgment results from the raters were reliable.

Finally, in order to investigate whether improvement in perception accuracy is a significant predictor of improvement in production accuracy, we adopted the perception accuracy data from the participants in Lee and Lyster (2016b) because they are the same as in the current study. Table 3 summarizes the participants' production and perception accuracy across groups and time for the trained words. Similarly, Table 4 displays their production and perception accuracy across groups and time for the untrained words.

Between-group analyses before the training and amount of feedback instances per group

To ensure that the five experimental groups were similar in production accuracy prior to the perception training, the participants were randomly assigned to each group. In addition, several sets of univariate analysis of variance (ANOVA) with an α level of 0.05 were conducted to detect any group differences in their accuracy before the training. The statistical assumptions for all the ANOVAs (e.g., Levene tests) were at first confirmed. The analyses confirmed that the groups were similar in their production accuracy for the pair /i/–/ɪ/ in the trained words, $F\ (4, 95) = 0.38$, $p = .825$, and in the untrained words, $F\ (4, 95) = 0.31$, $p = .870$, and for the pair /ɛ/–/æ/ in the trained words, $F\ (4, 95) = 0.44$, $p = .782$, and in the untrained words, $F\ (4, 95) = 0.19$, $p = .941$.

In order to ensure that one particular CF group did not receive more CF than the other CF groups, the number of CF occurrences was documented per group, revealing that the participants in the four CF groups received a similar amount of CF per training session (i.e., 384 trials): 99.2 ($SD = 35.2$), 105.2 ($SD = 52.7$), 91.6 ($SD = 45.2$), and 106.1 ($SD = 36.7$) CF occurrences per training session in the target, nontarget, combination, and wrong groups. An ANOVA with an

Table 3. *Mean (standard deviation) accuracy scores over time by group and target pair for trained words (n = 20 per group)*

	Pretest		Immediate Posttest		Delayed Posttest	
Group	/i/–/ɪ/	/ɛ/–/æ/	/i/–/ɪ/	/ɛ/–/æ/	/i/–/ɪ/	/ɛ/–/æ/
Production Accuracy (5-Point Scale)						
Target	1.56	1.69	2.49	3.13	2.25	2.31
	(1.03)	(1.02)	(0.54)	(0.18)	(0.16)	(0.76)
Nontarget	1.65	1.94	2.20	2.79	2.02	2.64
	(1.21)	(1.04)	(0.83)	(0.69)	(0.60)	(0.69)
Combination	1.55	1.92	1.54	2.04	1.57	2.30
	(1.10)	(0.85)	(0.35)	(0.12)	(0.31)	(0.54)
Wrong	1.64	1.65	1.76	2.06	1.18	2.00
	(0.99)	(0.80)	(0.57)	(0.34)	(0.33)	(0.39)
Control	1.28	1.69	1.22	1.25	1.04	1.16
	(1.06)	(1.04)	(0.50)	(0.61)	(0.35)	(0.54)
Perception Accuracy (100-Point Scale)						
Target	72.29	59.79	81.88	78.96	82.71	76.35
	(11.43)	(15.87)	(10.45)	(11.43)	(9.44)	(11.56)
Nontarget	70.52	59.06	79.27	77.40	79.48	74.69
	(15.18)	(17.39)	(14.11)	(10.86)	(13.42)	(12.27)
Combination	68.13	57.50	84.17	81.04	80.52	79.06
	(16.39)	(19.86)	(10.89)	(10.71)	(12.52)	(11.76)
Wrong	68.13	58.54	80.21	75.52	76.98	75.94
	(8.20)	(13.24)	(9.74)	(13.45)	(11.26)	(13.66)
Control	66.25	59.69	67.75	60.71	64.71	59.75
	(16.70)	(15.93)	(13.78)	(10.19)	(12.47)	(12.83)

Note: Production accuracy: 1 (*difficulty to understand*) ~ 5 (*easy to understand*).

α level of 0.05 revealed that there were no group differences in terms of the amount of CF, F (3, 76) = 0.48, p = .696, with the Levene test insignificant, p = .164.

RESULTS

Analysis 1: Gains in production accuracy

Mixed-design ANOVAs were conducted to investigate the gains in production accuracy after the perception training in the trained and untrained words, respectively. Each mixed-design ANOVA was designed with group as a between-subject independent variable (target, nontarget, combination, wrong, and control) and time as a within-subject independent variable (pretest, immediate posttest, and delayed posttest) with an α level of 0.05. Statistical assumptions such as

Table 4. *Mean (standard deviation) accuracy scores over time by group and target pair for untrained words (n = 20 per group)*

	Pretest		Immediate Posttest		Delayed Posttest	
Group	/i/–/ɪ/	/ɛ/–/æ/	/i/–/ɪ/	/ɛ/–/æ/	/i/–/ɪ/	/ɛ/–/æ/
Production Accuracy (5-Point Scale)						
Target	1.66	1.71	3.19	2.35	2.19	2.79
	(0.96)	(0.70)	(0.52)	(0.32)	(0.75)	(0.74)
Nontarget	1.75	1.82	1.35	1.15	1.26	1.29
	(1.24)	(1.00)	(0.86)	(0.59)	(0.49)	(0.30)
Combination	1.57	1.86	1.19	1.46	1.33	1.94
	(0.93)	(0.65)	(0.45)	(0.49)	(0.26)	(0.62)
Wrong	1.73	1.92	1.24	1.89	1.28	1.48
	(1.00)	(0.68)	(0.60)	(0.88)	(0.40)	(0.39)
Control	1.44	1.87	0.98	1.41	1.11	1.71
	(0.90)	(0.96)	(0.38)	(0.42)	(0.37)	(0.45)
Perception Accuracy (100-Point Scale)						
Target	68.96	56.92	80.62	69.79	80.63	68.54
	(13.28)	(9.96)	(12.04)	(13.31)	(10.92)	(10.34)
Nontarget	64.58	56.46	75.00	66.67	77.08	65.00
	(13.69)	(12.79)	(13.11)	(12.68)	(11.90)	(13.88)
Combination	66.25	58.13	78.12	67.29	79.79	66.46
	(15.82)	(15.02)	(12.53)	(11.96)	(12.11)	(12.28)
Wrong	68.12	57.46	80.42	68.75	79.17	64.38
	(8.25)	(9.58)	(11.40)	(13.89)	(10.73)	(11.67)
Control	63.45	53.88	62.79	56.75	68.54	59.46
	(11.54)	(10.36)	(17.48)	(10.06)	(18.36)	(13.54)

Note: Production accuracy: 1 (*difficulty to understand*) ~ 5 (*easy to understand*).

data normality, Levene tests, and Mauchly tests were verified before carrying out each mixed-designed ANOVA. Significant main effects of time (irrespective of the group variable) were not of primary interest and thus not further analyzed. Of primary interest were the significant Time × Group interaction effects, which entailed pairwise comparisons of each posttest and the pretest for each group using the Bonferroni correction. In the same vein, the effect sizes between each posttest and the pretest were calculated using the Cohen d (Cohen, 1988) and classified as small ($0.50 \leq d < 1.10$), medium ($1.10 \leq d < 1.60$), or large ($1.60 \leq d$) for within-group contrasts (Plonsky & Oswald, 2014).

As for the trained words, there was a significant main effect of time regarding the pair /i/–/ɪ/, F (2, 190) = 5.26, $p = .006$, and the pair /ɛ/–/æ/, F (2, 190) = 13.14, $p < .001$. The analyses also revealed a significant Time × Group interaction for the pair /i/–/ɪ/, F (8, 190) = 3.00, $p = .003$, and for the pair /ɛ/–/æ/, F (8, 190) = 6.89, $p < .001$.

Following up each interaction effect, pairwise comparisons found that the target group's production accuracy of the pair /i/–/ɪ/ at each posttest was significantly higher than that of the pair at the pretest (p = .001, d = 1.13 at the immediate posttest; p = .021, d = 0.93 at the delayed posttest). The target group's production accuracy of the pair /ɛ/–/æ/ was also significantly higher at each posttest than that of the pair at the pretest (p = .002 d = 1.96 at the immediate posttest; p = .015, d = 0.68 at the delayed posttest). Regarding the nontarget group, there was no significant improvement in its production accuracy of the pair /i/–/ɪ/ between each posttest and the pretest (p = .101, d = 0.54 at the immediate posttest; p = .418, d = 0.39 at the delayed posttest). However, its production accuracy of the pair /ɛ/–/æ/ at each posttest was significantly higher than that of the pair at the pretest (p = .002, d = 0.96 at the immediate posttest; p = .015, d = 0.78 at the delayed posttest). As for the other three groups, the analyses failed to reach significance between each posttest and the pretest regarding the two pairs /i/–/ɪ/ and /ɛ/–/æ/ and the effect sizes were small.

With respect to the untrained words, the analyses did not find a significant main effect of time regarding the pair /i/–/ɪ/, F (2, 190) = 2.33, p = .100, and the pair /ɛ/–/æ/, F (2, 190) = 2.71, p = .069. However, the analyses revealed a significant Time × Group interaction for the pair /i/–/ɪ/, F (8, 190) = 8.44, $p < .001$, and for the pair /ɛ/–/æ/, F (8, 190) = 6.42, $p < .001$.

Following up each interaction effect, pairwise comparisons regarding the target group revealed that its production accuracy of the pair /i/–/ɪ/ was significantly higher at the immediate posttest than at the pretest ($p < .001$, d = 1.97); however, its production accuracy at the delayed posttest was not (p = .098, d = 0.61). As for the pair /ɛ/–/æ/, the target group's production accuracy was significantly higher at each posttest compared to the pretest (p = .007, d = 1.18 at the immediate posttest; $p < .001$, d = 1.50 at the delayed posttest). Regarding the nontarget group, the only significant improvement was for the pair /ɛ/–/æ/ at the immediate posttest (p = .005; d = 0.82). No other comparisons for this group reached significance. Finally, with regard to the other three groups, there were no statistically significant changes in their production accuracy of either sound pair between each posttest and the pretest, and the effect sizes were small.

To summarize, concerning the target group, the production accuracy of both pairs in the trained words was significantly higher at both posttests with large to small effect sizes. The production accuracy of the pair /ɛ/–/æ/ in the untrained words was significantly higher at both posttests with medium effect sizes, whereas the production accuracy of the pair /i/–/ɪ/ in the untrained words was significantly higher only at the immediate posttest with a large effect size. As for the nontarget group, there were no significant changes in the production accuracy of the pair /i/–/ɪ/ in the trained and untrained words. As for the pair /ɛ/–/æ/, the production accuracy in the trained words was significantly higher at both posttests; however, the production accuracy in the untrained words was significantly higher at the immediate posttest only. All the effect sizes regarding the nontarget group were small. With respect to the combination, wrong, and control groups, there were no significant changes in the production accuracy of either pair in the trained and untrained words. Again, all the effect sizes pertaining to those groups were small.

Analysis 2: Relationship between improvement in perception accuracy and improvement in production accuracy

In order to delve into whether improvement in perception accuracy as a result of perception training is a significant predictor of improvement in production accuracy, regression analyses were conducted with an α level of 0.05 in the trained and untrained words each. To reiterate, we adopted the perception accuracy data from Lee and Lyster (2016b) to quantify improvement in participants' perception accuracy. In particular, the improvement in perception accuracy was computed by averaging two improvement scores; one was obtained by subtracting the scores of the immediate posttest from the scores of the pretest, and the other was prepared by subtracting the scores of the delayed posttest from the scores of the pretest. The improvement in production accuracy was also quantified in the same manner. Before conducting each regression analysis, the assumption of linearity was verified; homoscedasticity, normality, and independence were also examined. The regression analyses were conducted including all groups (i.e., all 100 Korean participants) except for the native baseline speakers.

As for the trained words, improvement in perception accuracy of the pair /i/–/ɪ/ was a significant predictor of improvement in its production accuracy ($R^2 = .05$, $b = 0.02$, $SE = 0.01$; $\beta = 0.22$, $t = 2.24$, $p = .028$). Similarly, improvement in perception accuracy of the pair /ɛ/–/æ/ was a significant predictor of improvement in its production accuracy ($R^2 = .05$, $b = 0.02$, $SE = 0.01$; $\beta = 0.22$, $t = 2.27$, $p = .025$). With respect to the untrained words, however, improvement in perception accuracy of the pair /i/–/ɪ/ was not a significant predictor of improvement in production accuracy of the pair ($R^2 = .01$, $b = 0.01$, $SE = 0.01$; $\beta = 0.08$, $t = 0.79$, $p = .431$). The regression analysis regarding the pair /ɛ/–/æ/ also failed to identify improvement in perception accuracy as a significant predictor of improvement in production accuracy ($R^2 = .01$, $b = 0.01$, $SE = 0.01$; $\beta = 0.09$, $t = 0.91$, $p = .365$).

DISCUSSION

Differential effects of L2 speech perception training on production accuracy

The first noteworthy finding in the current study is that, overall, participants improved their production accuracy of the target pairs with the aid of the L2 speech perception training. However, the extent to which the production accuracy of the L2 learners benefited from the perception training depended on CF type. That is, those who were in the target and nontarget groups improved their production accuracy after the perception training, whereas those in the combination, wrong, and control groups failed to do so. In particular, the results showed that the target condition is more effective than the nontarget condition with respect to inducing L2 learners to improve their production accuracy.

The perception training did not entail any production training insofar as the tasks did not require the participants to produce the target pairs. Nonetheless, the first author and research assistants observed a particular behavior during the training on the part of the participants in the target and nontarget groups but not those in the other groups. That is, most participants in the target and nontarget groups verbally responded to the CF on their perception errors. They were eager to utter the target form after each CF intervention. Hardison (2003) also revealed a similar finding: "I did note that learners in the AV (auditory-visual) training sessions imitated the lip movements of the talkers they were observing, although the training involved perception only, and their task was to circle the word from a minimal pair that they thought the talker on the screen was saying" (p. 516). Previous research on learner responses to CF revealed that, relative to other error types, CF on phonological errors led to the greatest amount of immediate learner uptake (Lyster, 1998; Sheen 2006), suggesting that this verbal behavior might be intrinsic to learner responses to CF on phonological errors.

Applying skill acquisition theory (DeKeyser, 1998, 2001; Lyster & Sato, 2013) to L2 phonological development, declarative knowledge refers to phonological representations of the target pairs encoded in memory, whereas procedural knowledge refers to abilities to actually produce language by accessing the L2 phonological representations. Anderson, Corbett, Koedinger, and Pelletier (1995) argued that the proceduralization of declarative knowledge is accomplished through practice and feedback. In this view, findings from the Lee and Lyster (2016b) study indicated that participants in the target and nontarget groups developed more targetlike phonological representations with the support provided by the perception training. In addition, in the present study, given that both types of CF ended up unintentionally inciting the participants to produce the target form, the target and nontarget conditions seem to have favored proceduralization. As a result, participants in the target and nontarget groups benefited from the perception training even in terms of production accuracy. In addition, the improvement in production was visible not only in the trained words but also to some extent, albeit somewhat less generalized, in the untrained words. This suggests that, within the target and nontarget groups, L2 learners' phonological representations of these two target pairs became more targetlike.

The target condition was expected to enhance the target pattern and thus to induce the participants to notice and become aware of the phonetic differences between the target and nontarget forms. As for the nontarget condition, it was expected to encourage the participants to consider the alternative by providing them with negative evidence that included the nontarget form; therefore, they were also predisposed to notice and become aware of the phonetic differences between the target and nontarget forms. As a result, such opportunities for noticing and awareness in the receptive mode helped L2 learners to reanalyze and restructure their L2 phonological representations into more targetlike representations. Moreover, considering that both types of CF created opportunities to produce the target form, the target and nontarget conditions inadvertently provided opportunities for production practice, which may have contributed to the proceduralization of these reanalyzed phonological representations.

With respect to production opportunities, the target condition would appear more conducive than the nontarget condition, owing to the nature of the CF types. Specifically, given that the target condition entailed provision of the target form immediately following a perception error, those in the target group were predisposed to verbally utter the target form. As for the nontarget condition, it was designed to draw the L2 learners' attention to the nontarget form so they would consider the alternative. Accordingly, it would seem less conducive for inciting participants to reproduce the form correctly, because the CF did not include the actual target form and instead pushed them to consider the alternative. In both cases, however, participants were observed uttering the target form, but these utterances were not expected and so were not quantified for further analysis.

Concerning the combination type, we found a mismatch between the perception and production outcomes. As reported in Lee and Lyster (2016b), the combination group showed the highest perception accuracy after the perception training. However, as for the production measurement in the present study, those in the combination condition did not show any significant improvement. By juxtaposing the target form with the nontarget form, the combination type of CF was designed to optimize awareness of phonetic differences (i.e., the psychoacoustic salience), and so the participants in the combination condition had robust opportunities for noticing and awareness as a means to achieve more targetlike phonological representations in declarative knowledge. However, compared to the participants in the target and nontarget conditions, those in the combination condition seemed not to engage as frequently in the verbal behavior (i.e., uttering the target form) and may thus have had fewer opportunities for proceduralization. This is of course an open question worthy of further investigation. In a similar vein, further studies are needed to answer the question of why the combination condition was not as conducive to inciting the L2 participants to produce the target form.

Compared to the three auditory CF types, the wrong type did not provide any linguistic information other than a statement that an error had been made. As a result, the wrong type was ineffective in helping the L2 learners to improve their perception accuracy. In addition, those in the control group participated in the same perception training; however, they were not able to confirm whether their L2 phonological representations were acceptable in the target language owing to the absence of CF (Lee & Lyster, 2016b). Given that accurate speech production might derive from targetlike L2 phonological representations (Flege, 1995; Flege et al., 2003), it might be the case that those in the wrong and control groups were not able to improve their production accuracy due to the absence of targetlike L2 phonological representations.

Finally, the current study found that the effects of the training including CF seemed to have a lasting effect on one pair but not the other. Previous studies (e.g., Flege et al., 1997) showed that Korean learners of English have more difficulty with /ɛ/–/æ/ than with /i/–/ɪ/. Thus, we speculate that there might be more room for improvement in the former pair compared to the latter.

Impacts of improvement in perception accuracy on production accuracy

As for the trained words, improvement in perception accuracy was a significant predictor of production accuracy. That is, the extent to which production accuracy improved depended on L2 learners' improvement in perception accuracy. This finding is compatible with previous empirical studies and the perception-first view in L2 phonological acquisition. In this view, the increased accuracy in L2 speech perception of the L2 participants in the present study served a bootstrapping function that enabled them to achieve more targetlike production. Nevertheless, considering that the regression analyses were significant for the trained words only, the impact of improvement in perception accuracy was not robust enough for us to draw a definitive conclusion, as was the case in other studies (e.g., Bradlow et al., 1997; Peperkamp & Bouchon, 2011), which also showed mixed results.

What draws our attention regarding the relationship between production and perception accuracy is that improvements in production and perception accuracy are not always parallel (Bradlow et al., 1997). In this sense, we argue that targetlike phonological representations might be necessary for targetlike production but do not necessarily guarantee targetlike production. That is, participants in the wrong and control conditions showed no improvement in perception accuracy (Lee & Lyster, 2016b) nor in production accuracy, owing arguably to the absence of targetlike L2 phonological representations. Participants in the combination condition improved their perception accuracy, and had thus encoded targetlike phonological representations, yet did not improve their production accuracy, owing again to the absence of opportunities for production practice and thus for proceduralization. Accordingly, for learners to achieve successful L2 speech production and perception, we recommend pedagogical interventions for L2 speech training that draw on skill acquisition theory by including opportunities for noticing, awareness, and practice, in addition to CF (Lyster, 2007; Ranta & Lyster, 2007).

Conclusion and future directions

Along with previous research, the current study found that the production accuracy of L2 learners benefits from L2 speech perception training without explicit production training. However, the extent of the benefits varies in accordance with CF type. In this respect, we argue that proper perception training be taken into consideration in L2 pronunciation instruction. We also recommend that CF treatments be carefully identified in L2 speech perception training and that their differential effects be accounted for in L2 speech acquisition.

In terms of production accuracy, the current study found that the target and nontarget types of CF were effective because they seemed to provide the L2 participants with opportunities to produce the target forms orally, which were in turn believed to expedite the proceduralization of L2 phonological representations. As for the combination type, the present study and the previous one by Lee and Lyster (2016b) revealed a mismatch between perception and production accuracy outcomes. While the combination type of CF was effective in improving the L2 learners' perception accuracy (Lee & Lyster, 2016b), this type of CF was not beneficial with respect to improving their production accuracy. On the one hand, we observed that the participants in the combination condition were less likely to produce target forms following CF, and we speculated that these were missed opportunities for proceduralization. On the other hand, we call for further studies to explore why they were less likely to do so.

During the perception training, the participants were allowed to listen to each stimulus repeatedly. Considering that excess input exposure might result in improvement in production accuracy regardless of CF type, it would be interesting to investigate whether similar results would obtain if the number of repetitions was controlled. In addition, the participants in the target and nontarget groups were eager to utter the target form after each CF intervention, whereas those in the other groups were not. Because we did not expect this phenomenon, we were not able to quantify the number of utterances per group. Accordingly, in order to solidify our argument regarding the roles of the CF types, we suggest that future studies be designed to quantify the number of utterances produced during L2 speech perception training with various CF types.

One might argue that L2 speech production training (e.g., explicit phonetic instruction) would result in similar findings and that speech production-perception mixed training would be ideal for both modalities. Comparison studies of the effects of L2 speech production-focused training, L2 speech perception-focused training, and L2 speech production-perception mixed training on both modalities would be thus of future interest. It would be also interesting to replicate the current study with different age groups, such as late versus early L2 learners (Saito, 2015) and children versus adults (Baker et al., 2008); with different learning contexts, such as laboratory versus classroom settings; and with different proficiency levels. The impact of length of instruction and residence in target language countries would be worth investigating as well (Saito & Brajot, 2013; Saito & Hanzawa, 2015). Finally, given that the current study adopted two particular vowel pairs in monosyllabic words and a production measurement at a controlled-speech level, we call for research studies with more fine-grained research designs (e.g., various measurement tasks) and several linguistic targets with various word/sentence environments to have a clearer understanding regarding the roles of various CF types in L2 phonological development. In particular, it would be valuable to scrutinize the influence of speech perception training and CF on L2 speech production of linguistic targets when they are embedded in multisyllabic words, when the words are located in various positions in a carrier sentence, and when L2 learners are prompted to produce the words at a spontaneous-level speech.

Nevertheless, along with Lee and Lyster (2016b), one finding is certain: the use of right or wrong as a primary means of providing CF in L2 speech perception training should be reconsidered.

ACKNOWLEDGMENTS

This research was supported by Social Sciences and Humanities Research Council of Canada Grant 410-2011-0671 (to R.L.). An earlier version of this article was presented at the 2016 annual meeting of the American Association for Applied Linguistics held in Orlando, Florida. We express our gratitude to all participants and research assistants Benjamin Gormley, Reginald Gooch, and Takashi Oba for helping us at various phases of this study. We also thank the anonymous reviewers and Associate Editor Isabelle Darcy for offering invaluable comments on earlier drafts.

REFERENCES

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences, 4*, 167–207.

Baker, W., Trofimovich, P., Flege, J. E., Mack, M., & Halter, R. (2008). Child-adult differences in second-language phonological learning: The role of cross-language similarity. *Language and Speech, 51*, 317–342.

Baker, W., Trofimovich, P., Mack, M., & Flege, J. E. (2002). The effect of perceived phonetic similarity on non-native sound learning by children and adults. In S. A. Fisch, B. Scarabela, & A.-H. Do (Eds.), *Proceedings of the 26th Boston University Conference on Language Development* (Vol. 26, pp. 36–47). Somerville, MA: Cascadilla Press.

Bent, T., Bradlow, A. R., & Smith, B. L. (2007). Phonemic errors in different word positions and their effects on intelligibility of non-native speech: All's well that begins well. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 331–347). Amsterdam: John Benjamins.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issue in cross-language research* (pp. 171–204). Timonium, MD: York Press.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). Amsterdam: John Benjamins.

Boersma, P., & Weenink, D. (2013). Praat: Doing phonetics by computer (Version 5.3.41) [Computer software]. Retrieved from http://www.praat.org

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

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

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

Colantoni, L., & Steele, J. (2008). Integrating articulatory constraints into models of second language phonological acquisition. *Applied Psycholinguistics, 29*, 489–534.

Davies, M. (2008). *The Corpus of Contemporary American English (COCA): 450 million words, 1990–present*. Retrieved from http://corpus.byu.edu/coca

de Jong, K., Hao, Y.-C., & Park, H. (2009). Evidence for featural units in the acquisition of speech production skills: Linguistic structure in foreign accent. *Journal of Phonetics, 37*, 357–373.

DeKeyser, R. M. (1998). Beyond focus on form: Cognitive perspectives on learning and practicing second language grammar. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 42–63). Cambridge: Cambridge University Press.

DeKeyser, R. M. (2001). Automaticity and automatization. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 125–151). Cambridge: Cambridge University Press.

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

Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition, 28*, 339–368.

Flege, J. E. (1995). Second-language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issue in cross-language research* (pp. 233–277). Timonium, MD: York Press.

Flege, J. E. (2002). Interactions between the native and second-language phonetic system. 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 Trier.

Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics, 25*, 437–470.

Flege, J. E., Schirru, C., & MacKay, I. R. A. (2003). Interaction between the native and second language phonetic subsystems. *Speech Communication, 40*, 467–491.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "L" and "R." *Neuropsychologia, 9*, 317–323.

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

Hardison, D. M. (2012). Second language speech perception: A cross-disciplinary perspective on challenges and accomplishments. In S. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 349–363). New York: Routledge.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America, 97*, 3099–3111.

Ingram, J. C., & Park, S. (1997). Cross-language vowel perception and production by Japanese and Korean learners of English. *Journal of Phonetics, 25*, 343–370.

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.

Kissling, E. M. (2014). What predicts the effectiveness of foreign-language pronunciation instruction? Investigating the role of perception and other individual differences. *Canadian Modern Language Review/La revue canadienne des langues vivantes, 70*, 532–558.

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. (2010). *A guide to doing statistics in second language research using SPSS*. New York: Routledge.

Lee, A. H., & Lyster, R. (2016a). The effects of corrective feedback on instructed L2 speech perception. *Studies in Second Language Acquisition, 38*, 35–64.

Lee, A. H., & Lyster, R. (2016b). Effects of different types of corrective feedback on receptive skills in a second language: A speech perception training study. *Language Learning, 66*, 1–25.

Lee, H. (1993). Korean. *Journal of the International Phonetic Association, 23*, 28–31.

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning, 60*, 309–365.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/: II. The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America, 94*, 1242–1255.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). San Diego, CA: Academic Press.

Lyster, R. (1998). Recasts, repetition, and ambiguity in L2 classroom discourse. *Studies in Second Language Acquisition, 20*, 51–81.

Lyster, R. (2007). *Learning and teaching languages through content: A counterbalanced approach*. Amsterdam: John Benjamins.

Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA. *Studies in Second Language Acquisition, 32*, 265–302.

Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback in second language classrooms. *Language Teaching, 46*, 1–40.

Lyster, R., & Sato, M. (2013). Skill acquisition theory and the role of practice in L2 development. In P. García Mayo, M. Gutierrez-Mangado, & M. Martínez Adrián (Eds.), *Contemporary approaches to second language acquisition* (pp. 71–92). Amsterdam: John Benjamins.

McDonough, K. (2007). Interactional feedback and the emergence of simple past activity verbs in L2 English. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A series of empirical studies* (pp. 323–338). Oxford: Oxford University Press.

Munro, M. J., & Derwing, T. M. (2008). Segmental acquisition in adult ESL learners: A longitudinal study of vowel production. *Language Learning, 58*, 479–502.

Neri, A., Cucchiarini, C., & Strik, H. (2006). Selecting segmental errors in non-native Dutch for optimal pronunciation training. *International Review of Applied Linguistics in Language Teaching, 44*, 357–404.

Nixon, R. (2012). *Learning PHP, MySQL, JavaScript, and CSS: A step-by-step guide to creating dynamic websites* (2nd ed.). Sebastopol, CA: O'Reilly Media.

Peperkamp, S., & Bouchon, C. (2011). The relation between perception and production in L2 phonological processing. In S. Trancoso (Chair), *12th Annual Conference of the International Speech Communication Association (Interspeech 2011)* (Vol. 1, pp. 168–171). Red Hook, NY: Curran Associates.

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.

Pulvermüller, F., & Schumann, J. H. (1994). Neurobiological mechanisms of language acquisition. *Language Learning, 44*, 681–734.

Ranta, L., & Lyster, R. (2007). A cognitive approach to improving immersion students' oral language abilities: The awareness-practice-feedback sequence. In R. M. DeKeyser (Ed.), *Practice in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 141–160). Cambridge: Cambridge University Press.

Saito, K. (2013). The acquisitional value of recasts in instructed second language speech learning: Teaching the perception and production of English /ɹ/ to adult Japanese learners. *Language Learning, 63*, 499–529.

Saito, K. (2015). The role of age of acquisition in late second language oral proficiency attainment. *Studies in Second Language Acquisition, 37*, 713–743.

Saito, K., & Brajot, F. (2013). Scrutinizing the role of length of residence and age of acquisition in the interlanguage pronunciation development of English /ɹ/ by late Japanese bilinguals. *Bilingualism: Language and Cognition, 16*, 847–863.

Saito, K., & Hanzawa, K. (2015). Developing second language oral ability in foreign language classrooms: The role of the length and focus of instruction and individual differences. *Applied Psycholinguistics*. Advance online publication.

Sheen, Y. (2006). Exploring the relationship between characteristics of recasts and learner uptake. *Language Teaching Research, 10*, 361–392.

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

Strange, W. (2006). Second-language speech perception: The modification of automatic selective perceptual routines. *Journal of the Acoustical Society of America, 120*, 3137.

Strange, W. (2007). Selective perception, perceptual modes, and automaticity in first- and second-language processing. *Journal of the Acoustical Society of America, 122*, 2970.

Strange, W., & Shafer, V. L. (2008). Speech perception in second language learners: The re-education of selective perception. In J. G. H. Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 153–191). Amsterdam: John Benjamins.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. M. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Rowley, MA: Newbury House.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 125–144). Oxford: Oxford University Press.

Thomson, R. I. (2011). Computer assisted pronunciation training: Targeting second language vowel perception improves pronunciation. *CALICO Journal, 28*, 744–765.

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

Uther, M., Knoll, M. A., & Burnham, D. (2007). Do you speak E-NG-LI-SH? A comparison of foreigner- and infant-directed speech. *Speech Communication, 49*, 2–7.

Wang, X., & Munro, M. J. (2004). Computer-based training for learning English vowel contrasts. *System, 32*, 539–552.

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *Journal of the Acoustical Society of America, 113*, 1033–1043.

Watkins, K., & Paus, T. (2004). Modulation of motor excitability during speech perception: The role of Broca's area. *Journal of Cognitive Neuroscience, 16*, 978–987.

Yang, B. (1996). A comparative study of American English and Korean vowels produced by male and female speakers. *Journal of Phonetics, 24*, 245–261.