Effects of Form-Focused Instruction and Corrective Feedback on L2 Pronunciation Development of /ɹ/ by Japanese Learners of English

Kazuya Saito

McGill University

Roy Lyster

McGill University

Sixty-five Japanese learners of English participated in the current study, which investigated the acquisitional value of form-focused instruction (FFI) with and without corrective feedback (CF) on learners' pronunciation development. All students received a 4-hr FFI treatment designed to encourage them to notice and practice the target feature of English /ɹ/ in meaningful discourse, except those in the control group ($n = 11$), who received comparable instruction but without FFI on English /ɹ/. During FFI, the instructors provided CF only to students in the FFI + CF group ($n = 29$) by recasting their mispronunciation or unclear pronunciation of /ɹ/, whereas no CF was provided to those in the FFI-only group ($n = 25$). Acoustic analyses were conducted on frequency values of the third formant (F3) of English /ɹ/ tokens elicited via pretest and posttest measures targeting familiar items and a generalizability test targeting unfamiliar items. The results showed that: (a) F3 values of the FFI + CF group significantly declined after the intervention, not only at a controlled-speech level but also a spontaneous-speech level, regardless of following vowel contexts; (b) change in F3 values of the FFI-only

group and the control group was not statistically significant; and (c) the generalizability of FFI to novel tokens remained unclear.

Keywords form-focused instruction; pronunciation instruction; corrective feedback; L2 phonology; English /ɹ/; English as a foreign language

Research into form-focused instruction (FFI) and corrective feedback (CF) has focused almost exclusively on morphosyntactic targets, in spite of calls for research into the roles for form-focused instruction in phonological development and suggestions that the latter might be especially amenable to phonological recasts. The current study extends the scope of FFI research by investigating the effects of FFI with and without CF on the acquisition of English-specific /ɹ/ by Japanese learners of English.

Form-focused instruction techniques draw attention to target language features that learners would otherwise not use or even notice in communicatively oriented classroom input (Spada, 1997). Unlike more traditional language instruction, FFI entails "a set of psycholinguistically motivated pedagogic options" (Ellis, 2001, p. 12) that are considered most effective when implemented in communicative contexts, to ensure that learners will be able to transfer what they learn in the classroom to communicative interaction outside the classroom. Among effective FFI activities identified by researchers, those of direct relevance to the present study include (a) structured input (i.e., learners process linguistic form in the input for meaning without being pressured to produce output; VanPatten, 2004), (b) typographically enhanced input (i.e., target structures are highlighted by means of intonational stress or visual changes such as italics or boldface to induce learners to notice the forms in oral and written second language [L2] input; Han, Park, & Combs, 2008), and (c) focused tasks (i.e., communicative activities designed to create obligatory contexts that elicit learners' use of a specific linguistic feature in production; Ellis, 2001, 2006). Such preplanned FFI interventions are often referred to as "proactive," whereas FFI interventions that include provision of CF in a seemingly less planned fashion during online interaction are considered "reactive" (Doughty & Williams, 1998). In the present study, we compare the relative effectiveness of proactive FFI interventions with and without a reactive component that includes provision of CF following learner errors.

Many observational and experimental studies of CF have investigated the effectiveness of recasts, which are defined as a teacher's reformulation of all or part of a student's utterance, minus the error. Although recasts have been identified as the most frequent type of CF used by teachers in a wide range

of instructional settings, some studies have pointed out that recasts of gram-
matical errors are potentially ambiguous for classroom learners accustomed
to focusing more on communication, because such recasts might appear to
be identical or alternative ways of saying the same thing in order to confirm
message comprehensibility or veracity (Ellis & Sheen, 2006; Lyster, 1998b,
2007; Lyster & Saito, 2010a; Nicholas, Lightbown, & Spada, 2001). In con-
trast, many descriptive studies have suggested that recasts might be relatively
effective for L2 phonological development compared to other domains such
as L2 morphosyntax, arguably because learners tend to perceive the corrective
force of teachers' recasts of pronunciation errors (Carpenter, Jeon, MacGregor,
& Mackey, 2006; Ellis, Basturkmen, & Loewen, 2001; Lyster, 1998a; Lyster
& Saito, 2010b; Mackey, Gass, & McDonough, 2000; Sheen, 2006). To the
best of our knowledge, there exist no experimental studies with a pretest and
posttest design that investigate the acquisitional value of FFI with and without
CF for L2 phonological development, especially during a set of meaningful FFI
tasks. With a general lack of intervention research in the area of pronunciation
teaching and an absence of research specifically investigating CF effectiveness
in phonological development, the time is ripe to explore this new direction.

Pronunciation Teaching

In the 1970s, pronunciation teaching was considered a priority in L2 classrooms
by proponents of the audio-lingual approach, who emphasized mastery of
nativelike pronunciation (especially phonemic contrasts) through the use of
minimal-pair drills and imitation of appropriate models (for a comprehensive
review of various approaches to pronunciation teaching, see Celce-Murcia,
Brinton, & Goodwin, 1996). The nativeness assumptions in the audio-lingual
approach were, however, not well supported by L2 speech research evidence,
which has convincingly shown that (a) L2 speech is typically foreign-accented,
mainly due to the interaction between the learners' age and the first language
(L1) (e.g., Flege, Munro, & MacKay, 1995; Piske, MacKay, & Flege, 2001),
and (b) very few adult learners achieve nativelike pronunciation in their L2
(Ioup, Boustagi, El Tigi, & Moselle, 1994; Moyer, 1999).

   The inevitability of foreign accent led many researchers as well as practi-
tioners to consider pronunciation as an unteachable subject and, consequently,
as Celce-Murcia et al. (1996) and Levis (2005) pointed out, to completely ig-
nore pronunciation teaching in their L2 instructional syllabi. Yet, there now
exists a revived interest in pronunciation teaching, based on the premise that
the ultimate goal of L2 speech learning is to achieve not only accurate but

also fluent usage of "intelligible" pronunciation for the purpose of success-ful L2 communication. Instead of aiming to eliminate pronunciation errors to foster accent-free speech, researchers who support this view stress that instruc-tion should focus only on aspects of pronunciation that influence intelligibility and comprehensibility in ways that make L2 communication more successful (Derwing & Munro, 2005; Field, 2005; Levis, 2005; Setter & Jenkins, 2005).

Some studies have examined the effects of explicit instruction on segmental aspects of L2 pronunciation via phonetic transcriptions and repetition practice (e.g., Rivers & Temperley, 1978) as well as suprasegmental aspects of L2 pro-nunciation through computer-mediated visual feedback (e.g., for a review of audiovisual training studies, see Hardison, 2010). Others have investigated the effects of intensive perceptual training on L2 speech perception (e.g., Iverson, Hazan, & Bannister, 2005; Lively, Logan, & Pisoni, 1993; Logan, Lively, & Pisoni, 1991) and its impact on L2 speech production (Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999; Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Hardison, 2003; Lambacher, Martens, Kakehi, Marasinghe, & Molholt, 2005). Yet, the relevance of these studies to real classroom settings can only be indirect at best, because they focused on the isolated teaching of "difficult" sound rules under strict laboratory conditions in which variables such as intensity and consistency of instruction were well controlled and the length of instruction on only one phonological target, in some cases, lasted for many hours (e.g., 15–22.5 hr in Bradlow et al., 1997, 1999; 11.5 hr in Hardison, 2003).

Although few in number, other studies have further investigated the ef-fects of pronunciation instruction by conducting quasi-experimental studies in actual classrooms (e.g., Couper, 2006; Elliott, 1997; Macdonald, Yule, & Powers, 1994; Neri, Mich, Gerosa, & Giuliani, 2008; for summaries of class-room studies of pronunciation teaching, see Derwing & Munro, 2005; Setter & Jenkins, 2005). One of the most often cited studies is Derwing, Munro, and Wiebe (1998), who investigated how a 10-week instructional treatment targeting either segmentals or suprasegmentals differentially impacted not only accentedness but also comprehensibility of learners' pronunciation. The gains made by students depended on the method of evaluation. With respect to ac-centedness, although students in the segmental and suprasegmental groups alike showed significant improvement in a sentence-reading task, they showed no significant gains in a picture-description task. With respect to comprehen-sibility, only students in the suprasegmental group demonstrated significant improvement in the picture-description task. The researchers concluded that pronunciation teaching studies need to take into account not only the target

of instruction (e.g., segmentals, suprasegmentals) and outcome measures (e.g., sentence reading, picture description) but also aspects of improvement (e.g., accentedness, comprehensibility, intelligibility).

In terms of type of instruction, most pronunciation teaching studies have depended on the exclusive use of explicit instruction on phonetic transcriptions followed by decontextualized practice such as mechanical drills and repetition, and their focus was apparently on forms rather than meaning. As DeKeyser (1998) pointed out, "It is rather uncontroversial that pronunciation is relatively immune to all but the most intensive formS-focused treatments [i.e., decontextualized language-focused methods]" (p. 43). One of the reasons for the dominance of focus-on-formS practice in pronunciation teaching could be that pronunciation requires not only metalinguistic knowledge (i.e., pronunciation rules) but also physical action (i.e., motor activities); that is, L2 learners need to develop abilities to manipulate articulatory organs properly to produce correct L2 sounds (see Flege, 2003, for a discussion of peculiarities of L2 speech production compared to other L2 skills).

Some L2 researchers, however, are doubtful of the effects of decontextualized instruction on learners' communicative competence, and they call for further research incorporating more psycholinguistically motivated instructional options in pronunciation teaching that would include tasks which are not only extensive and repetitive but also authentically communicative (Celce-Murcia et al., 1996; Pennington, 1996; Segalowitz, 2003; Trofimovich & Gatobonton, 2006). In her review of instructed second language acquisition (SLA) studies, Doughty (2003) noted:

> given the completely decontextualized nature of explicit focus on forms, this type of instruction promotes a mode of learning that is arguably unrelated to SLA, instructed or otherwise, in that the outcome is merely the accumulation of metalinguistic knowledge about language. (p. 271)

In fact, the results of previous pronunciation teaching studies have shown that the effects of explicit instruction followed by decontextualized practice on learners' improvement at a spontaneous speech level (measured by picture-description tasks) was rather discouraging (e.g., Elliott, 1997; Macdonald et al., 1994); other studies did not even test learners' extemporaneous speech production at all (e.g., Neri et al., 2008). Notably, Derwing et al.'s (1998) study did yield some positive results (i.e., participants receiving suprasegmental-based instruction showed improvement in picture-description tasks). Their goal was to compare a focus on suprasegmentals and segmentals via a mixture of various teaching methods (e.g., pronunciation lessons and some meaningful activities

such as group presentations) rather than isolate and test the effectiveness of a specific teaching method on one phonological target. In order to conduct a fine-grained analysis of how instruction facilitates L2 pronunciation development, intervention studies are called for that carefully spell out (a) what suprasegmental and segmental aspects of pronunciation are specifically targeted and (b) in what way and for how long the intervention is implemented to teach each of the target rules (Saito, 2011a).

Second-language grammar studies have convincingly shown that psycholinguistically motivated instructional treatments integrating form and meaning (FFI, focus on form) are more effective than (a) decontextualized teaching methods (grammar-translation methods, focus-on-forms) and (b) mere exposure to the target language (naturalistic approach, focus on meaning) (Doughty, 2003; Ellis, 2001, 2006; Lyster, 2007; Spada, 1997). Given the overall effectiveness of FFI on grammatical development (see meta-analyses by Norris & Ortega, 2000; Spada & Tomita, 2010), we consider it timely to explore the feasibility of FFI tasks that target pronunciation in meaning-oriented contexts and to assess their impact on L2 pronunciation development.

Measuring Pronunciation Development

In their research synthesis of 49 instructed SLA studies published between 1980 and 1998, Norris and Ortega (2000) excluded pronunciation teaching studies, noting that research designs of pronunciation teaching studies need to be critically reconsidered in order to provide any educational implications for L2 classrooms (see also Spada & Tomita, 2010). In this section, we specifically address the challenge of designing reliable outcome measures to assess the impact of instruction on learners' intelligible pronunciation development at a fine-grained level. We will point out some problems in this respect in previous studies and propose a solution that we adopted to conduct the present study.

Most of the previous pronunciation teaching studies cited ealier adopted human rating methods—that is, asking native-speaker (NS) listeners to rate nonnative-speaker (NNS) speech samples. The validity of the method has been substantially confirmed as the golden standard to measure the quality of sentence-level speech samples in a holistic manner (i.e., listeners generally agree with their rating judgment on perceived accents with high interrater reliability; see Derwing & Munro, 2009). Whereas listening tasks can be the best fit for cross-sectional data, they might not, however, be the most appropriate outcome measure for time-series data (as in intervention studies) for which a number of students need to be recruited and tested via various kinds of tasks

(eliciting both controlled- and spontaneous-speech production) at several times (i.e., pretest and posttest sessions). Human rating methods arguably take time that risks causing listener fatigue and limits the number of participants and speech tokens within one study (e.g., Derwing et al., 1998, for 6 hr of listening; Elliot, 1997, for 6 hr of listening in total).

As a reliable way to examine only acoustic properties of ample speech samples such as frequency values of formants, intensity, and duration at an individual word level, L2 phonology researchers tend to draw on an acoustic analysis; this kind of robust analysis enables researchers to measure change in the acoustic properties of L2 sounds between pretest and posttest sessions (see Saito, 2011). One could argue, however, that it is unclear how such changes in acoustic properties can actually impact NS listeners' comprehension of L2 speech production (which is arguably the ultimate goal of pronunciation teaching). Thus, as optimal outcome measures for pronunciation teaching studies of this kind, we propose one possible framework—a combination of an acoustic analysis and human rating method; that is, NS listeners are first recruited to rate a small subset of speech data randomly selected from the original data pool in order to find significant acoustic properties that positively influence NS listeners' rating scores. Second, acoustic analyses are conducted on the entire dataset with a focus on these significant acoustic variables. The assumption here is that, given that some acoustic properties are significantly correlated to NS listeners' comprehension, changes in such crucial acoustic properties will either immediately or eventually enhance overall intelligibility of L2 speech production.

English /ɹ/

Form-focused instruction in the current study targets one of the most well-researched cases of L2 speech sound learning—the acquisition of the English /ɹ/ by Japanese learners of English (for a review, see Bradlow, 2008). Due to the lack of any corresponding approximant sounds in the Japanese phonetic system (Japanese has only two approximants, /w/ and /j/), Japanese learners of English are predisposed to substituting the Japanese alveolar lateral flap /ɺ/ for the English /ɹ/ and /l/ and thereby neutralize the contrast in their production, even after many years of residence in English-speaking countries (Larson-Hall, 2006). They also judge both English /ɹ/ and /l/ perceptually as poor exemplars of the Japanese flap /ɺ/ (Best & Strange, 1992; Iverson et al., 2003).

Importantly, the current study specifically focused on the acquisition of English /ɹ/ rather than /l/. A number of studies have demonstrated that English /ɹ/ is acquired more easily by Japanese learners, both in terms of perception and

production, in contrast to English /l/ (Aoyama, Flege, Guion, Akahane-Yamada, & Yamada, 2004; Bradlow et al., 1997; see also Flege, 2003, for theoretical arguments). The authors of these studies argue that the difference between /l/ and /ɹ/ is more perceptually salient than /l/ and /l/ to Japanese learners of English, who, in turn, have greater facility in learning the articulatory configuration for /ɹ/ (as distinct from the Japanese flap /l/). NS listeners' perceptions of Japanese learners' speech sound production revealed a clear and significant improvement in learners' production of the sound /ɹ/ after both naturalistic and structured exposure to the language—in contrast, once again, to the sound /l/. Furthermore, recent perception research has revealed that in Japanese learners of English, the phonemic representation of the Japanese flap overlaps with that for English /l/, whereas /ɹ/ is more easily distinguished from /l/ (Hattori & Iverson, 2009; for a review, see Bradlow, 2008).

Building upon these findings, we argue that Japanese learners of English may actually benefit from a focus on English /ɹ/ as an initial step in developing clear perceptual and articulatory representations of the /ɹ/-/l/ contrast, despite a noticeable accent that may otherwise affect their intelligibility. Doing so could provide an efficient means of establishing an important sound contrast in English and help learners improve overall communicative success in their L2.

Pedagogical Importance of English /ɹ/

Some researchers emphasized the relative importance of suprasegmentals in successful L2 communication (e.g., Derwing et al., 1998; cf. Jenkins, 2000, 2002). From a pedagogical perspective, however, English /ɹ/ can be considered as a top priority especially for Japanese learners of English to improve overall intelligibility of their L2 speech production. In preparation for the proposed research, the first author administered an "expert judgment" questionnaire (see Ellis, 2006; Robinson, 1996) to a sample of 120 teachers of English in Japan comprising both NSs and NNSs of English, asking them to rank 25 pronunciation problems, which included a number of segmental problems (e.g., /ð/, /θ/, /f/, /v/, /æ/) as well as suprasegmental problems (e.g., lexical and sentential stress, speech rate, fluency). Results corroborated previous findings that English /ɹ/ is considered the most crucial teaching/learning target, owing to its potential to affect the intelligible pronunciation of Japanese learners of English (Saito, 2011b). Furthermore, several pronunciation specialists also argue that the English /ɹ/ and /l/ contrast, which has a relatively large number of frequently occurring minimal pairs, needs to be considered as one of the top teaching/learning targets not only for Japanese learners of English

but also other English as a second or foreign language (ESL/EFL) students worldwide, because of its high functional load on listeners' comprehension (see Munro & Derwing, 2006). Thus, examining this well-researched as well as pedagogically important topic is expected to yield benefits that reach well beyond Japanese learners of English, insofar as the findings may be generalizable to other ESL/EFL students worldwide (i.e., EFL learners in East Asia; ESL learners in North America) as well as different pronunciation features (i.e., segmentals vs. suprasegmentals).

Acoustic Properties of English /ɹ/
Previous L2 phonology studies have examined which acoustic properties (e.g., frequency values of the first formant [F1], second formant [F2], third formant [F3], and transitional duration of F3) determine NS listeners' categorical perceptions of English /ɹ/ and /l/ contrasts based on natural speech tokens (e.g., Espy-Wilson, 1992; Flege, Takagi, & Mann, 1995) as well as synthesized samples (e.g., Hattori & Iverson, 2009; Iverson & Kuhl, 1996; Iverson et al., 2003). Their findings generally suggest that the acoustic difference between /ɹ/ and /l/ depends primarily on the frequency values of F3[1]; that is, NS listeners tend to perceive the sound as /ɹ/ when its F3 dips below 2,000 Hz and as /l/ when its F3 exceeds 2,400 Hz or more (see also Ladefoged, 2003). In order to see if these findings (i.e., F3 as a primary phonetic cue) could be applicable to the current study, in which speech tokens were naturally produced across different tasks with various ensuing vowel contexts, we included a rating session during which five NS listeners evaluated a small subset of the data. This strategy allowed us to examine the extent to which the relevant acoustic properties varied according to task type and ensuing vowel contexts.

Current Study

We conducted a quasi-experimental study with a pretest and posttest design to investigate the effects on L2 pronunciation of FFI with and without CF (i.e., FFI-only and FFI + CF). Based on previous L2 grammar studies, FFI was operationalized in the current study in a set of production tasks designed to develop participants' argumentative skills in English while drawing their attention to the target forms through: (a) structured input; (b) typographically enhanced input; and (c) focused tasks. CF involved pronunciation-focused recasts (i.e., the partial, declarative type described by Sheen, 2006). In addition, we developed outcome measures through a combination of human ratings and acoustic analysis, and we tested their

validity by measuring the impact of instruction on the learners' L2 performance in various phonetic contexts not only at a controlled-speech level but also at a spontaneous-speech level. The research questions to be addressed in the current study are as follows:

1. Which acoustic properties of /ɹ/ affect NS listeners' judgments and how do these properties vary relative to task type and to the backness and height of ensuing vowels?
2. Does FFI lead to improvement in learners' pronunciation of /ɹ/ and do the effects of FFI vary according to whether or not learners also receive CF?

Method

The current study involved 2 experienced ESL teachers, 5 experienced NS listeners, and 65 adult intermediate-level Japanese learners of English in Montreal, Canada. The study comprised two phases. In the instructional phase, 65 learners were randomly divided into three groups (i.e., FFI-only group, FFI + CF group, and control group) and received 4 hr of meaning-based instruction about argumentative skills taught by the two ESL teachers. In addition, the FFI-only and FFI + CF groups received instruction on the English /ɹ/ sound. In the assessment phase, a rating session first took place in which the five NS listeners were asked to rate a small subset of speech samples randomly selected from the original data pool of speech tokens elicited at the pretest sessions. The goal was to ascertain which acoustic properties in English /ɹ/ (i.e., F1, F2, F3, transition duration) affected NS listeners' perceptional patterns according to task types and following vowel contexts. Subsequently, based on the results of a multiple regression analysis, an acoustic analysis was conducted on significant speech properties of all of the speech data produced in the pretest and posttest sessions. The aim was to see whether the learners showed any significant improvement on three tasks designed to tap different types of L2 speech production (i.e., controlled and spontaneous performance) as well as two following vowel contexts (English /ɹ/ followed by high/low vowels and front/back vowels).

Participants
Students
For the purpose of student recruitment, the first author created ads that advertised an opportunity to participate in a 4-hr free course on argumentative skills in English, specifying the proficiency levels required for participation (e.g., 450–700 for TOEIC scores, 50–80 for TOEFL iBT scores); our purpose was

to recruit intermediate Japanese learners of English based on the assumption that they would still have problems producing /ɹ/. The ad was posted on several community Web sites specifically for Japanese people studying abroad, with hardcopy versions also distributed at many language institutes in Montreal. Interested participants contacted the first author through email or by phone, and they set up a date for their first interview and pretest sessions. The recruitment continued until the number of participants reached the maximum number, which had been set in advance at 72. However, because four participants did not complete the instructional treatment nor attended the posttest sessions and three others were considered too advanced based on the pretest scores, there were a total of 65 participants included in the final analysis (age: $M = 29.7$, $SD = 6.9$).

During the first interview, a majority of participants reported that they attended either university-level English-speaking schools or private language institutes and had many opportunities to use English academically and socially on a daily basis. All of them had studied English for more than 10 years since their entrance to seventh grade in junior high school in Japan. Although most of the students had just arrived in Montreal, their length of residence (LOR) in Canada varied widely from 1 month to 13 years (LOR: $M = 15.5$ months, $SD = 31.8$ months). TOEIC scores were reported by 27 of the participants ($M = 577$, $SD = 168.12$). As a group, therefore, the participants were considered intermediate ESL learners. After the interviews, the 65 students were first randomly divided into 12 classes (6 students per class; cf. e.g., Sheen, 2006), and then these classes were designated as one of three groups: (a) FFI-only group (five classes, $n = 29$), (b) FFI + CF group (five classes, $n = 25$), and (c) control group ($n = 11$). Table 1 provides the details of the 65 participants' information according to the three groups.

Instructors

The two ESL teachers were both female and had several years of L2 teaching experience, including ESL/EFL instruction, and had worked at private language institutes in Montreal prior to the time of the project. Both were certified teachers with undergraduate degrees in L2 teaching and both were completing M.A. degrees in L2 education. They were selected on the basis of their professional and academic backgrounds and on their willingness and availability to participate. Both teachers followed training sessions, which will be described below. One instructor taught the first six classes (three FFI-only classes and three FFI + CF classes), and the other taught the other six classes (two FFI-only classes, two FFI + CF classes, and two control classes).

Table 1 Participant information by group

	FFI + CF group (n = 29)		FFI-only group (n = 25)		Control group (n = 11)	
	Mean	SD	Mean	SD	Mean	SD
Age	29.2	6.0	29.7	7.1	30.9	9.1
LOR (months)	18.5	31.5	11.5	28.7	18.0	40.8
Age of arrival	28.4	4.9	28.8	7.2	29.3	6.4
TOEIC	550.4	135.6	630.4	192.0	425.0	176.7

Note. LOR = length of residence, TOEIC = Test of English for Internal Communication, FFI = form-focused instruction, CF = corrective feedback. Gender distribution was as follows: 4 males and 25 females in the FFI + CF group, 5 males and 20 females in the FFI-only group, and 2 males and 9 females in the Control group.

Listeners

Five NSs of English (three males, two females) were recruited as NS listeners to rate the quality of 100 speech tokens randomly selected from the data pool of 1,430 speech tokens produced at the pretest sessions. The five NS listeners participating in the present study were selected on the basis of two crucial variables: their L1 variety of English and their familiarity with Japanese-accented English speech. Thus, although all five were undergraduate students studying at an English-speaking university in Montreal at the time of the study, they were all originally from the United States and spoke northeastern American English as their L1. All of them took Japanese classes and reported having frequent contact with Japanese learners of English in Montreal and being familiar with Japanese-accented English speech, including mispronunciation and unclear pronunciation of English /ɹ/. We thus considered them as "experienced" listeners (for the influence of accent familiarity on NS listeners' intelligibility judgment, see Bradlow & Bent, 2008; Kennedy & Trofimovich, 2008).

Procedure

Students in the experimental groups received 4 hr of FFI, which was designed to encourage them to notice and practice the target feature in the context of meaning-oriented instruction. Although those in the control group received comparable instruction (English argumentative skills), the target pronunciation feature of their FFI was different (English vowel sounds). The instructors gave CF only to students in the FFI + CF group by recasting their mispronunciation or unclear pronunciation of /ɹ/, whereas no CF was directed at those in the FFI-only group.

The three conditions implemented in the study (FFI-only, FFI + CF, and control) are described in full detail in the online supporting information Appendix S1, which also describes the pedagogical materials and the training the two teachers underwent prior to delivering each lesson. To recapitulate here, the FFI treatments on the pronunciation targets were integrated in the English argumentative skills lessons, which contained 38 minimally paired words (including near minimally paired words; see Table 2). Among these 38 words, English /ɹ/ appeared in various positions: 25 occurrences in word-initial positions, 3 in word-medial position, and 10 in consonant clusters. All words appeared frequently in various tasks, and they were italicized and highlighted in red so that the learners could notice the target feature during meaning-oriented tasks.

The FFI + CF group did the same activities involving the pronunciation targets as the FFI group, plus CF. CF was operationalized as pronunciation-focused recasts, referred to as "partial recasts" by Sheen (2006); that is, the

Table 2 Target minimally paired words

/ɹ/ in word-initial position		/ɹ/ in word-medial position	/ɹ/ in consonant cluster
*race	*road	arrive	bread
*rain	*rock	correct	crab
*ram	rocket	pirate	crime
rate	Rome		crowds
*read	*room		fries
*red	round		fruit
reef	*row		grass
rent	*rule		green
*right	*run		free
rice	*Ryan		pray
ring	*wrong		
rink	wrap		
river			

Note. The asterisk indicates the words included in the pretests/posttests.

instructors were asked to recast only one word in which an error occurred, with falling intonation and without adding any additional meaning (for examples, see the online supporting information Appendix S1). Arguably, recasts of this kind are so explicit that their corrective function is likely to be quite obvious. Finally, the participants in the control group received 4 hr of comparable instruction also on the topic of "developing English argumentative skills" but without form-focused instruction on English /r/ and with no exposure to the 38 target words.

Each class consisted of four 1-hr lessons and took place twice a week, finishing within 2 weeks (1 hr × 2 lessons per week × 2 weeks = 4 hr). The entire project (12 classes) took place over 7 months between March and September. All of the classes were conducted in a classroom located on the campus of an English-speaking university in Montreal. All instructional treatments were videotaped (4 hr of instruction × 12 classes = 48 hr) while the first author sat at the back of the classroom to ensure that the consistency of the instruction was maintained within groups by the two instructors. Two weeks after the 4 hr of instruction, the students individually completed posttests as well as final interviews.[2] Figure 1 summarizes the design of the study and the procedures followed.

It is important to note here that, although the students were made aware of the main purpose of the project (i.e., English argumentative skills) at the first
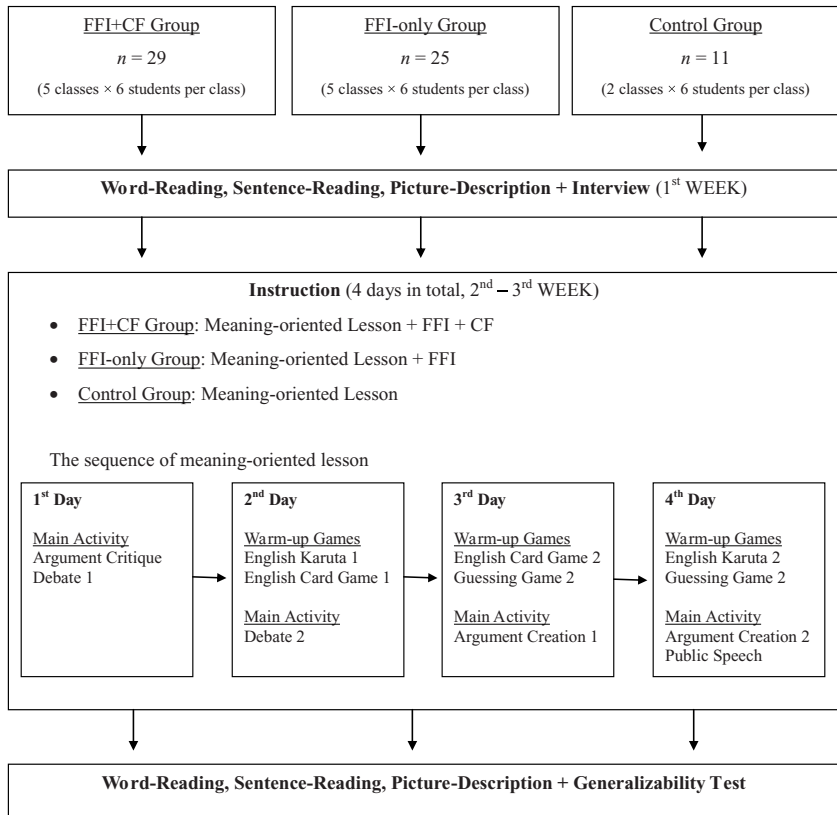
Figure 1 Summary of the procedure.

interview, the other focus of the project—the pronunciation-related FFI part—was purposely not explained to them at all until the introspective interview was conducted at the end point of the experiment. This is because one of the purposes of the study was to investigate whether and in what ways FFI with or without CF induces students to notice the target feature and practice it during meaningful discourse without any explicit explanations.

Measures

In order to examine the impact of instruction on learners' oral production of English /ɹ/ from various perspectives, three tests targeting familiar items were administered as pretest and posttest measures before and after the instructional treatment, and a fourth test targeting unfamiliar items was administered on

only one occasion after the treatment. The testing sessions were completed in a quiet room in one-on-one meetings with the first author, a NS of Japanese. All communication and instruction about the procedure was always done in Japanese so that the learners never heard any model pronunciation of the target words from the researcher.[3] Their speech tokens were carefully recorded by means of the computer speech recognition software, Praat (Boersma & Weenik, 2009), at a 44.1-kHz sampling rate and a 16-bit resolution. A unidirectional microphone was used (DM-20SL) and all of the recordings were stored on the hard drive of a TOSHIBA Satellite U400 laptop computer.

Familiar Items

Three tests targeting 14 of the 38 words that had appeared in the instructional treatment were used to evaluate the learners' improvement in their pronunciation of English /ɹ/ between pretests and posttests: a word-reading task, a sentence-reading task, and a timed picture-description task. The word- and sentence-reading tasks were designed to measure controlled performance (i.e., explicit knowledge), whereas the picture-description task was designed to elicit learners' spontaneous speech production (i.e., implicit knowledge). The same tasks were used at both pretests and posttests (hence the importance of a control group to assess possible test-retest effects).

Word Reading

Learners read 25 individual words, which included 10 target words and 15 distracters. All target words were consonant-vowel-consonant (CVC) singleton tokens beginning with /ɹ/ (i.e., *read, room, root, rule, red, race, rough, row, ram*, *right*).

Sentence Reading

In addition to four distractor sentences, participants read the following five sentences, in which eight target words were embedded:

  He will read my paper by the time I arrive there.
  She left her red bicycle on the side of the road.
  The race was cancelled because of the rain.
  I can correct all wrong sentences tonight.
  Ryan does not like to run in the snow.

All target items were CVC singleton tokens (i.e., *read, red, road, race, rain, wrong*, *run*) with one exception of CVVC (i.e., *Ryan*).[4]

Timed Picture Description

Learners were asked to describe four pictures designed to elicit four CVC singleton tokens (i.e., *read, rain, rock, road*) as well as four distracter pictures that did not include any target words. Adjacent to each picture were three word cues to prompt learners to use the target word while describing the scene. For example, a picture of a table left on a driveway in the rain was accompanied by three word cues (i.e., *table, driveway, rain*) and was used to elicit the target word *rain* in the participant's description. Participants were given only 5 seconds to prepare before being prompted by the researcher to begin their description so that they were required to perform the task under time pressure while their main focus was on meaning (see Ellis, 2005).

Whereas the 38 minimally paired words targeted by the instruction included English /ɹ/ in various positions (word-initial, word-medial, consonant clusters), /ɹ/ occurred only in word-initial positions in the 14 test items. All but one were CVC singletons (see asterisked words in Table 2). This decision was made because Japanese learners of English tend to have an especially difficult time in their perception and production of English /ɹ/ in word-initial positions compared to word-final positions (Bradlow et al., 1997; Goto, 1971; Lively et al., 1993; Logan et al., 1992; Mochizuki, 1981; Sheldon & Strange, 1982) and because the assumption in the current study was that measuring the learners' performance of English /ɹ/ on relatively difficult positions (word-initial) would reveal their current proficiency levels.[5]

Twenty-two word-initial singleton tokens were elicited per participant at pretest and posttest sessions, respectively (out of 14 words, 8 were tested twice but in different tasks; the others appeared only once). The results were analyzed separately based on: (a) three different tasks ($n = 10$ for word reading, $n = 8$ for sentence reading, and $n = 4$ for picture description); (b) following vowel backness ($n = 10$ for singletons with front vowels, $n = 2$ for singletons with mid vowels, $n = 10$ for singletons with back vowels); and (c) following vowel height ($n = 5$ for singletons with high vowels, $n = 14$ for singletons with mid vowels, $n = 3$ for singletons with low vowels).

According to previous research, Japanese learners of English display more difficulty in perceiving English /ɹ/ following rounded vowels (i.e., /o/, /u/) than any other context (e.g., Hardison, 2003; Ingram & Park, 1998; Sheldon & Strange, 1982). The current study further pursued this topic by investigating differential effects of training based on various following vowel contexts: (a) backness (front, mid, back vowels) and (b) height (high, mid, low).

Unfamiliar Items

At the posttest only, a generalizability test was administered to 60 participants to investigate whether they could generalize their newly acquired knowledge of English /ɹ/ to unfamiliar items (5 students could not take the test for personal reasons). Participants were asked to read a list of four non-minimally-paired words that had not appeared during the instructional treatment (i.e., *real, roll, rumor, regular*) along with four distracter items. The results were analyzed to detect any between-group differences according to vowel contexts (i.e., backness and height of vowels immediately following /r/).

Rating and Judgment Sessions

Given the relatively large number of participants in the current study ($n = 65$), asking NS listeners to rate a huge number of speech samples ($n = 3,100$ tokens) could have caused listeners' fatigue, which, in turn, would have forced us to limit either the number of test tokens or the number of participants. As an alternative that would better allow us to achieve a precise description of the nature of L2 speech development under instructed conditions, we adopted a combination of human rating methods and acoustic analysis in the present study.

One hundred speech tokens were randomly chosen from the original data pool of the learners' performance in the pretest sessions and then presented to the five NS listeners to rate. To select the 100 speech tokens, 20 learners were first chosen randomly (4 participants from the control group, 7 from the FFI-only group, and 9 from the FFI + CF group) and each of them contributed 5 words (2 from the word-reading task; 2 from the sentence-reading task; 1 from the picture-description task). These words were carefully edited from the original sound files by the first author by means of the speech analysis software (Praat). In the case of all 100 tokens—especially those extracted from the sentence-reading and picture-description tasks—every effort was made to isolate the tokens by following the procedure described next.

First, a spectrographic representation of each word was displayed on the computer screen using Praat. Then the researcher listened to each token several times and tried to locate the beginning and end of the word without including any trace of the preceding and following sounds. As a reliable clue, the starting point of English /ɹ/ was identified by the end point of the gradually falling transition of F3. The F3 of the preceding vowel and consonant sounds tend to gradually fall, because the F3 for English /ɹ/ is relatively low (e.g., Bradlow, 2008; Hattori & Iverson, 2009; Iverson et al., 2003; see for details of acoustic properties of English /ɹ/ Espy-Wilson, 1992; Flege, Takagi, & Mann, 1995;

Ladefoged, 2003). At the onset of the target word, the researcher placed a cursor, which he then moved to the end of the word in order to cut and paste it into a separate sound file. In this manner, all 100 tokens were prepared and put on one data CD to be used in the rating session.

The rating sessions took place individually with each listener in a quiet room and lasted about 1 hr. First, after a briefing about the purpose of the current study, each NS listener completed a training session with five speech samples. After the training, the NS listener was presented 100 speech tokens in a randomized order and asked to rate them on a 9-point scale, with 1 as "very good English /ɹ/" and 9 as "very poor English /ɹ/," using the whole scale as much as possible. Following this was an intelligibility judgment task in which the listener was asked to decide whether the speech token sounded like the English /ɹ/ or the English /l/, again following a brief training session. The NS listeners were allowed to listen to each speech token as many times as they wanted until they felt confident about their ratings and judgments.[6] In addition, the first author was always next to the listeners during their sessions so that he could answer any questions they had regarding the rating procedure and ensure consistency in the rating and judgment procedures.

Acoustic Analysis

To conduct the acoustic analysis of English /ɹ/ across various contexts, we adopted the procedures used by Flege, Takagi, & Mann (1995) in their analysis of the /ɹ/ and /l/ contrast produced in various tasks by Japanese learners of English. Accordingly, the first author measured F1, F2, and F3 values in hertz (Hz) and transitional duration of F3 in milliseconds (ms) for all speech tokens. He also categorized speech tokens based on task types (word reading, sentence reading, picture description) and also according to the height (high, mid, low) and frontness/backness (front, mid, back) of subsequent vowels.

As for F1, F2, and F3, as described earlier, the beginnings of the English /ɹ/ were first carefully identified by the end point of falling F3, and then a cursor was put on the location where energy was clear for all three formants and F3 was starting to rise (i.e., the end of the steady state). The procedure is visually summarized in Figure 2. As for the transitional duration of F3, the researcher first put a cursor on the starting point of rising F3 and moved it to the beginning of the following vowels (measured in milliseconds).

Finally, to determine which speech and nonspeech factors predicted listeners' rating and judgment patterns, two multiple regression analyses were conducted on the rating and judgment scores with two types of listening tasks as dependent variables (i.e., 9-point scale and intelligibility judgment) and
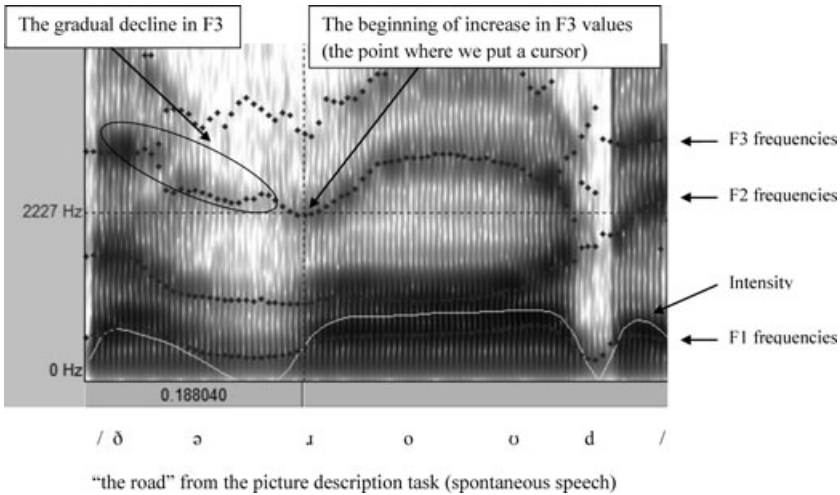
Figure 2 Procedure for the acoustic analysis for English /ɹ/.

seven predictors as independent variables (i.e., F1, F2, F3 transition duration, subsequent vowel height, subsequent vowel frontness/backness, and task types). After identifying crucial acoustic properties as significant predictors of listeners' 9-point scores and intelligibility scores alike, we moved on to the acoustic analyses of the remaining speech tokens ($n = 3,100$) in order to examine the learners' improvement in conjunction with relevant acoustic properties between the pretest and posttest sessions.

For comparison purposes, we asked the five NS listeners to do the same four tasks (i.e., word reading, sentence reading, picture description, and generalizability test) so that we could obtain baseline data conveying NS production of the four acoustic properties (F1, F2, F3, and transition duration). They contributed a total of 130 speech tokens (5 NS talkers × 26 tokens = 130 tokens).

Results

We first present (a) which acoustic properties were significantly correlated to listeners' rating scores and then further examine (b) to what extent the learners improved in their pronunciation of familiar and unfamiliar tokens.

Analysis of Acoustic Properties and Listener Ratings
Because the interclass correlation among the five NS listeners on the 9-point rating scale proved to be strongly correlated ($r = .78$, $p < .0001$), their scores

were averaged for each speech token. To identify which factors predicted their rating scores, a multiple regression analysis was conducted on the listeners' average scores as dependent variables and seven predictor factors as independent variables: (a) F1, (b) F2, (c) F3, (d) transition duration, (e) subsequent vowel height, (f) subsequent vowel frontness/backness, and (g) task types. Although the original model proved significant, $F(7, 92) = 12.593, p < .0001$, collinearity statistics found that both F2 and F3 exhibited relatively high variation inflation factor (VIF) values of more than 2.0 (VIF = 2.26 and 2.01, respectively). In fact, these two factors were highly correlated with one another ($r = .64, p < .0001$). Following previous research findings whereby F3 is a primary phonetic cue for NS listeners, only the F3 factor was taken into account; the second multiple regression analysis was conducted with six predictor factors excluding F2: (a) F1, (b) F3, (c) transition duration, (d) subsequent vowel height, (e) subsequent vowel frontness/backness, and (f) task type. This second model proved to be significant, $F(6, 93), p < .0001$, revealing only F3 as a significant predictor factor ($t = 6.269, p < .0001$).

With respect to listeners' intelligibility judgment, the interclass correlation was also strongly correlated ($r = .72 \; p < .0001$). Therefore, a multiple regression analysis was again conducted on their average scores as a dependent variable and seven predictor factors as independent variables (F1, F2, F3, transition duration, vowel height, vowel frontness/backness, task types), revealing collinearity problems based on the VIF values of F2 and F3 (VIF = 2.26 and 2.01, respectively). Consequently, a second multiple regression analysis was conducted with six independent variables excluding F2: (a) F1, (b) F3, (c) transition duration, (d) vowel height, (e) vowel frontness/backness, and (f) task type. The results showed that the model was significant, $F(6, 93) = 9.775 \; p < .0001$, revealing only F3 as a significant predictor factor ($t = 6.09, p < .0001$).

In their goodness of English ratings on a 9-point scale, the listeners perceived F3 values around 2,230 Hz as very good English /ɹ/ ($1 < x < 3$), F3 values around 2,363 Hz as less ambiguous English /ɹ/ ($4 < x < 6$), and F3 values around 2,780 Hz as very poor English /ɹ/. Similarly, in their intelligibility judgments (i.e., the speech token sounds like /ɹ/ or like /l/), all five listeners judged sounds with F3 values around 2,270 Hz as definitely /ɹ/ and sounds with F3 values at around 2,800 as definitely /l/. Sounds with F3 values around 2,400 and 2,600 were judged as either English /ɹ/ or English /l/. Details of these results are displayed in Table 3.

Following our finding that F3 frequency values were the only significant predictor of NS listeners' perception of English /ɹ/, we proceeded to conduct acoustic analyses on the rest of the entire data focusing on F3 values, in order

Table 3 Average F3 values of the 100 speech tokens rated by 5 NS listeners

	n	M	SD
Goodness of English (9-point scale)			
Very good /ɹ/ (1 < x < 3)	32	2,230 Hz	354 Hz
Neutral /ɹ/ (4 < x < 6)	30	2,363 Hz	370 Hz
Very poor /ɹ/ (7 < x < 9)	38	2,780 Hz	383 Hz
Intelligibility judgment			
Judged as /ɹ/ by 5 listeners	50	2,270 Hz	382 Hz
Judged as /ɹ/ by 4 listeners	10	2,416 Hz	359 Hz
Judged as /ɹ/ by 3 listeners	5	2,611 Hz	295 Hz
Judged as /ɹ/ by 2 listeners	4	2,765 Hz	199 Hz
Judged as /ɹ/ by 1 listener	9	2,546 Hz	342 Hz
Judged as /ɹ/ by no listeners	22	2,804 Hz	248 Hz

to investigate whether and to what degree the learners' F3 values changed as a result of the FFI treatment. In order to interpret these results, we first considered any reduction in F3 values as improvement (positive impact on NS listeners' perception) and then examined the degree of improvement by measuring the amount of decline between pretest and posttest sessions (e.g., a transition of F3 values from 2,600 to 2,200 Hz suggests that the learners' unclear pronunciation of English /ɹ/ becomes more intelligible).

Word-Reading, Sentence-Reading, and Picture-Description Tasks
The 65 learners each produced 44 tokens (22 at each teaching session) for a grand total of 2,860 tokens. First, as for the pretest scores ($n = 22$) and posttest scores ($n = 22$), we averaged across each participant's F3 values according to (a) task types (word reading, sentence reading, and timed picture description), (b) following vowel backness (singletons with front vowels, mid vowels, and back vowels) and (c) following vowel height (singletons with high vowels, mid vowels, and low vowels), respectively. Second, for each contextual factor (task type, following vowel backness, and height), we conducted separate three-factor ANOVAs in order to identify statistically significant differences between pretest and posttest sessions (within-group comparison) and among the three groups at the posttest sessions (between-group comparison). The alpha level was set at a $p < .05$ level for all statistical analyses. Cohen's d was also calculated in order to measure the magnitude of instructional effectiveness between two contrast groups of means.[7]

Task Type

To assess effects of instruction on learners' speech production of English /ɹ/ both at controlled- and spontaneous-speech levels, a three-factor ANOVA was conducted: Group (FFI + CF, FFI-only, control) × Task (word reading, sentence reading, timed picture description) × Time (pretests/posttests). The ANOVA results revealed significant effects for the overall Group × Task × Time interaction, $F(4, 124) = 2.635, p = .0373$. A simple main effect of Time was found significant only for the FFI + CF group at all contexts: (a) word reading ($M = 2,511 \rightarrow 2,339$ Hz), $F(1, 186) = 15.647, p < .00001, d = 0.59$; (b) sentence reading ($M = 2,542 \rightarrow 2,342$ Hz), $F(1, 186) = 21.030, p < .00001, d = 0.76$; and (c) timed picture description ($M = 2,622 \rightarrow 2,385$ Hz), $F(1, 186) = 29.676, p < .00001, d = 0.81$. In addition, a simple main effect for Group was significant for the timed picture-description task at posttest sessions. A Tukey test was conducted as a post hoc analysis with its alpha level set at a $p < .05$, revealing that only the FFI + CF group ($M = 2,385$ Hz) exhibited significantly lower F3 values than the control group ($M = 2,695$ Hz) on the timed picture-description task with large effects ($d = 1.14$). Their F3 values generally declined from 2,600–2,700 Hz (confusing English /ɹ/) to 2,300 Hz (less ambiguous English /ɹ/), which indicates significant effects of FFI with CF on their L2 speech production of English /ɹ/. Interestingly, the ANOVA results also found main effects for Task, $F(2, 124) = 12.245, p < .000001$. According to a Tukey test, F3 values significantly differed at a $p < .001$ level in the following order: word reading ($M = 2,451$ Hz) < sentence reading ($M = 2,495$ Hz) < timed picture description ($M = 2,547.45$ Hz).

According to the production baseline data that had also been collected from the five NSs, the mean of F3 values was 1,648 Hz ($SD = 212$ Hz) for the word-reading task, 1,677 Hz ($SD = 226$ Hz) for the sentence-reading task, and 1,692 Hz ($SD = 221$ Hz) for the timed picture-description task. Note that the NS F3 values (around 1,700 Hz) differed substantially from those produced by Japanese learners of English (around 2,500 Hz). Not surprisingly, different from those of Japanese learners of English, the F3 values of the NS baseline data did not significantly differ across the three tasks, $F(2, 8) = 3.101, p = .100$ (i.e., NS talkers produced English /ɹ/ with little variance regardless of different tasks).

Following Vowel Backness

To examine how instruction effectiveness varied according to types of tokens (English /ɹ/ following front, mid, and back vowels), a three-factor ANOVA was conducted: Group (FFI + CF, FFI-only, control) × Backness (singletons

with front vowels, mid vowels, and back vowels) × Time (pretests/posttests). The ANOVA results revealed significant effects for the overall Group × Time interaction, $F(2, 62) = 7.337$, $p = .00014$. A simple main effect for Time was found significant only for the FFI + CF group ($M = 2,538 \rightarrow 2,321$ Hz), $F(1, 62) = 31.090$, $p < .00001$, $d = 0.63$. The decline in their F3 values can be interpreted as their improvement from confusing English /r/ to less ambiguous English /ɹ/. Furthermore, an overall main effect for Backness was also identified as significant, $F(2, 124) = 37.495$, $p = .00014$. According to a Tukey test, F3 values significantly differed between tokens with back and mid vowels ($M = 2,426$ Hz and $2,427$ Hz, respectively) and front vowels ($M = 2,554$ Hz).

The results of the NS baseline data showed that the mean of their F3 values was 1,717 Hz (199 Hz) for singletons with front vowels, 1,640 Hz (189 Hz) for singletons with mid vowels, and 1,661 Hz (174 Hz) for singletons with back vowels. Although their F3 values (around 1,700 Hz) were substantially different from those of Japanese learners of English (around 2,500 Hz), they did not significantly differ according to following vowel backness, $F(2, 8) = 2.801$, $p = 0.134$.

Following Vowel Height
To examine how instruction effectiveness varied according to types of tokens (English /ɹ/ following high, mid and low vowels), a three-factor ANOVA was conducted: Group (FFI + CF, FFI-only, control) × Height (singletons with high vowels, mid vowels, and low vowels) × Time (pretests/posttests). The ANOVA results revealed significant effects for the overall Group × Time interaction, $F(2, 62) = 8.472$, $p = .0006$. A simple main effect of Time was significant only for the FFI + CF group ($M = 2,554 \rightarrow 2,355$ Hz), $F(1, 62) = 30.596$, $p < .00001$, $d = 0.81$. A simple main effect of Group also proved significant at the time of posttests, $F(2, 124) = 3.444$, $p < .05$. A Tukey test further showed that the FFI + CF group ($M = 2,329$ Hz) outperformed the control group ($M = 2,529$ Hz) with large effects ($d = 0.93$). The change in their F3 values ($2,500–2,600$ Hz $\rightarrow 2,300$ Hz) could be considered as evidence that their unclear pronunciation of English /r/ became less ambiguous. An overall main effect for Height was identified as significant, $F(2, 124) = 20.581$, $p < .00001$. According to a Tukey test, F3 values significantly differed according to following vowel height in the following manner: mid vowels ($M = 2,455$ Hz) < low vowels ($M = 2,496$ Hz) < high vowels ($M = 2,558$ Hz).

The mean of the NS F3 values was 1,682 Hz ($SD = 183$ Hz) for singletons with high vowels, 1,630 Hz ($SD = 234$ Hz) for singletons with mid vowels, and 1,646 Hz ($SD = 184$ Hz) for singletons with low vowels. As was the case with

the other contexts, the mean of their F3 values (around 1,700 Hz) was greatly different from that of Japanese learners of English (around 2,500 Hz) and did not significantly differ according to following vowel height, $F = (2, 8), p = 0.438$. The results of pretests and posttests are summarized in Table 4.

Generalizability Test
In order to assess participants' pronunciation of English /ɪ/ in words that had not appeared during the instructional treatment, we conducted a separate two-way ANOVA on the group means of the F3 values yielded by the generalizability test according to two contextual factors: (a) following vowel backness (singletons with front and back vowels) and (b) following vowel height (singletons with high and mid vowels).

Following Vowel Backness
Although the two-way ANOVA (Group × Backness) found an overall main effect for Backness, $F(1, 57) = 36.704, p < .00001$, it was not statistically significant for the overall Group effect, $F(2, 57) = 0.034, p = .362$. Singletons with back vowels ($M = 2,247$ Hz) exhibited significantly lower F3 values than those with front vowels ($M = 2,401$ Hz), but there existed no significant difference among F3 values of the FFI + CF group ($M = 2,289$ Hz), the FFI-only group ($M = 2,313$ Hz), and the control group ($M = 2,434$ Hz). In spite of the lack of statistical significance, effect size analyses showed that both the FFI-only and FFI + CF groups began to exhibit small-to-medium effects compared to the control group ($d = 0.45$ for the FFI-only group, $d = 0.72$ for the FFI + CF group) in their pronunciation of English /ɪ/ following front vowels (where F3 values are relatively high). The results of the NS baseline data exhibited no significant difference in mean F3 values for front vowels (1,754 Hz) and back vowels (1,635 Hz), $F(1, 4) = 3.989, p = 0.1185$.

Following Vowel Height
The two-way ANOVA (Group × Height) identified an overall main effect for Height, $F(1, 57) = 19.330, p < .00001$, but not for Group, $F(2, 57) = 1.035, p < .3617$; that is, the F3 values of all three groups were significantly different between high vowels ($M = 2,392$ Hz) and mid vowels ($M = 2,250$ Hz). Nevertheless, small-to-medium effects were found for both the FFI-only group ($d = 0.33$ for high vowels and 0.38 for mid vowels) and the FFI + CF group ($d = 0.51$ for high vowels and 0.47 for mid vowels). The results of the NS baseline data exhibited no significant difference in the mean F3 values for high

Table 4 Summary of results of pretests and post tests targeting familiar items

1. Group × Task × Time
   Contextual factors (F3 values)
   - Word reading ($M = 2{,}451$ Hz) < sentence reading ($M = 2{,}495$ Hz) < timed picture description ($M = 2{,}695$ Hz)

   Between-group difference (F3 values)
   - For timed picture description, FFI + CF ($M = 2{,}385$ Hz) < control ($M = 2{,}695$ Hz)

   Within-group differences
   - For word reading: FFI + CF at posttests ($M = 2{,}339$ Hz) < FFI + CF at pretests ($M = 2{,}511$ Hz)
   - For sentence reading: FFI + CF at posttests ($M = 2{,}342$ Hz) < FFI + CF at pretests ($M = 2{,}542$ Hz)
   - For timed picture description: FFI + CF at posttests ($M = 2{,}385$ Hz) < FFI + CF at pretests ($M = 2{,}622$ Hz)

   NS baseline
   - No significant difference between tasks ($M = 1\ 648$ Hz for word reading; $M = 1\ 677$ Hz for sentence reading; $M = 1\ 692$ Hz for timed picture description)

2. Group × Backness × Time
   Contextual factors (F3 values)
   - Low and mid vowels ($M = 2{,}426$ Hz and $2{,}427$ Hz, respectively) < front vowels ($M = 2{,}554$ Hz)

   Within-group difference (F3 values)
   - For all contexts: FFI + CF at posttests ($M = 2{,}355$ Hz) < FFI + CF at pretests ($M = 2{,}544$ Hz)

   NS baseline
   - No significant difference between vowel backness ($M = 1{,}717$ Hz for front vowels; $M = 1{,}640$ Hz for mid vowels; $M = 1{,}661$ Hz for back vowels)

3. Group × Height × Time
   Contextual factors (F3 values)
   - Mid vowels ($M = 2{,}455$ Hz) < low vowels ($M = 2{,}496$ Hz) < high vowels ($M = 2{,}558$ Hz)

   Between-group difference (F3 values)
   - For all contexts: FFI + CF ($M = 2{,}329$) < control ($M = 2{,}529$)

   Within-group difference (F3 values)
   - For all contexts: FFI + CF at posttest ($M = 2{,}355$ Hz) < FFI + CF at pretests ($M = 2{,}554$ Hz)

   NS baseline
   - No significant difference between vowel height ($M = 1{,}682$ Hz for high vowels; $M = 1{,}630$ Hz for mid vowels; $M = 1{,}646$ Hz for low vowels)

Table 5 Summary of results of generalizability test targeting unfamiliar items

---

1. Group × Backness × Time
    Contextual factors (F3 values)
        • Low vowels ($M = 2{,}247$ Hz) < front vowels ($M = 2{,}401$ Hz)
    Effect size analysis (compared to control group)
        • FFI-only (small-to-medium effects)
        • FFI + CF (small-to-medium effects)
    NS baseline
        • No significant difference between vowel backness ($M = 1{,}635$ Hz for back
          vowels; $M = 1{,}754$ Hz for front vowels)
2. Group × Height × Time
    Contextual factors (F3 values)
        • Mid vowels ($M = 2{,}250$ Hz) < high vowels ($M = 2{,}392$ Hz)
    Effect size analysis (compared to control group)
        • FFI-only (small-to-medium effects)
        • FFI + CF (small-to-medium effects)
    NS baseline
        • No significant difference between vowel height ($M = 1{,}658$ Hz for high vowels;
          $M = 1{,}631$ Hz for mid vowels)

---

vowels (1,658 Hz) and for mid vowels (1,631 Hz), $F(1, 4) = 0.241, p = 0.6495$.
The results of the generalizability test are summarized in Table 5.

Personal Interview

At the onset and end point of the project, the learners were interviewed by the
first author in face-to-face meetings. Among several questions asked that were
not the focus of this article, one was highly relevant to the current study and
will thus be reported here. After they finished the posttest sessions, the learners
were asked what they had learned the most from the 4-hr instruction. Out of
the 65 learners who completed the project, 63 learners reported "English argu-
mentative skills" such as debating and public speaking skills as their *primary*
concerns and "the importance of an English /ɹ/ and /l/ contrast" as their *sec-
ondary* concern. Only two learners (both from the FFI + CF group) reported
that their focus was always on form (i.e., English /ɹ/) because the content of the
lesson (English argumentative skills) was beyond their English proficiency.

Discussion

The first research question asked which acoustic properties of /ɹ/ affect NS
listeners' judgments and whether they vary according to task type, following

vowel backness and following vowel height. F3 values for English /ɹ/ were iden-
tified as the most crucial speech properties but with some variance according
to task type and differences in vowel backness and height. This finding resulted
from outcome measures that combined human rating and acoustic analysis.
The second research question asked whether FFI improves learners' pronun-
ciation of /ɹ/ and whether its effects increase through provision of CF. Results
revealed that learners receiving FFI without CF did not show any significant
change in their F3 values, whereas those receiving FFI in conjunction with CF
generally decreased their F3 values from 2,600–2,500 Hz (confusing English
/ɹ/) to 2,200–2,300 Hz (less ambiguous English /ɹ/), which, in turn, suggests a
significant improvement in their pronunciation of English /ɹ/ as a result of the
FFI treatment with CF. A detailed discussion of the results is now presented.

Acoustic properties of English /ɹ/
Based on 100 speech tokens randomly selected from the 1,430 speech tokens
produced at the pretest sessions, a multiple regression analysis confirmed that,
among various independent variables, only F3 values were a significant predic-
tor factor for NS listeners' rating scores (for similar results, see Espy-Wilson,
1992; Flege, Takagi, & Mann, 1995; Hattori & Iverson, 2009; Iverson et al.,
2003; Ladefoged, 2003). Post hoc analyses revealed several patterns regarding
the relationship between NS listeners' perception of English /ɹ/ and F3 values:
(a) Speech tokens with F3 values around 2,200–2,300 Hz tended to be con-
sidered as both "good-enough" exemplars of English /ɹ/ and definitely English
/ɹ/ rather than English /l/; (b) those with F3 values around 2,400–2,600 Hz
proved to be confusing to NS listeners, who judged them as either English /ɹ/ or
English /l/; and (c) speech tokens with F3 values above 2,600 Hz were judged
either as poor exemplars of English /ɹ/ or definitely as English /l/. Interestingly,
learners' F3 values varied significantly according to task type, following vowel
backness and following vowel height, whereas those of the NS talkers did not
show any significant variance associated with contextual factors.

Important to emphasize is that the listeners' judgment in the current study
(i.e., speech tokens with F3 values around 2,200–2,300 Hz are close enough
to English /ɹ/) is relatively lenient rather than strict, because their judgment
of good exemplars (F3 values around 2,200–2,300 Hz) is still significantly
different from the NS baseline data with F3 values at around 1600–1700 Hz.
In other words, not only did the current study identify F3 as a crucial speech
property that significantly influences listeners' judgment of English /ɹ/, but it
also set realistic goals for L2 learners in terms of intelligible pronunciation of

English /ɹ/ (F3 around 2,200–2,300 Hz) rather than nativelike pronunciation of English /ɹ/ (F3 around 1,600–1,700).

Task Type

The results showed that the Japanese learners of English in the current study tended to produce significantly higher F3 values (less nativelike production of English /ɹ/) when cognitive demands increased according to the three different task types: word reading ($M = 2,451$ Hz) < sentence reading ($M = 2,495$ Hz) < timed picture description ($M = 2,695$ Hz). Although the learners exhibited the highest F3 values (confusing English /ɹ/) in the timed picture-description task, one could argue that the latter might not be the most appropriate way to tap learners' spontaneous speech production of English /ɹ/ (i.e., the learners were asked to describe pictures, but they were provided orthographic representations of target words as cues they could then read). Thus, we call for future research that further develops robust elicitation techniques appropriate for assessing spontaneous speech production (see Piske et al., 2001).

Following Vowel Backness and Height

The learners had difficulties producing English /ɹ/ following high front vowels /i/, demonstrating relatively high F3 values for front vowels ($M = 2,554$ Hz) and high vowels ($M = 2,558$ Hz). Intriguingly, previous research found that it is most difficult for Japanese learners of English to *perceive* the English /ɹ/ and /l/ contrast following round vowels such as /u/ and /o/ (e.g., Hardison, 2003). Taken together, it might be the case that learners' processing of the target language is different in production (i.e., English /ɹ/ following high front vowels is the most difficult instance for production) and perception (English /ɹ/ following round back vowels is the most difficult instance for perception), which might, in turn, contribute to the asymmetry in production and perception abilities of Japanese learners of English for the /ɹ/ and /l/ contrast (i.e., some learners produce English /ɹ/ better than perceive it, and vice versa; Goto, 1971; Mochizuki, 1981; Sheldon & Strange, 1982). Again, more research is needed to further pursue this topic.

Other Phonetic Cues

The findings about F3 as a primary phonetic cue for the English /ɹ/ and /l/ contrast in the current study support previous research findings in experimental phonetics and L2 phonology studies (Espy-Wilson, 1992; Flege, Takagi, & Mann, 1995; Hattori & Iverson, 2009; Iverson et al., 2003; Ladefoged, 2003).

Noteworthy is that several studies found not only F3 but also other speech properties as significant predictors for NS listeners' perceptional patterns for the English /ɹ/ and /l/ contrast (e.g., Polka & Strange, 1985, for F1 transition length). Although we acknowledge the possibility of considering other speech properties such as F1 transition duration as independent variables, the types of speech tokens and the listening procedure adopted in the current study were substantially different from previous studies, which might have influenced our finding that only F3 values were significant predictors. First, compared to synthetic speech samples typically used in previous studies, the current study used natural speech tokens, which were less systematically controlled, and so the listeners might have failed to capture the subtle influence of so-called secondary phonetic cues such as F1 transition. Second, we carefully recruited only experienced NS listeners who had familiarity with Japanese accented English /ɹ/ and /l/, which might have influenced their tendency to judge even relatively high F3 as English /ɹ/. In any case, we still need to wait for more L2 speech research to be done, further investigating how NS listeners use various phonetic cues (e.g., F1, F2, F3, transition duration) to discern the English /ɹ/ and /l/ contrast in the context of natural speech tokens as well as synthesized speech.

FFI With and Without CF

In the current study, whereas the frequency and salience of the target form was enhanced through FFI that drew attention to the target pronunciation form of English /ɹ/ in the context of meaning-oriented tasks, participants in the FFI + CF group were also given pronunciation-focused recasts by their teachers in response to their mispronunciation and unclear pronunciation. The effects of instruction will be discussed based on the results of tests targeting familiar items and unfamiliar items, respectively.

When FFI was combined with CF, the learners' F3 values generally changed on familiar items from 2,500–2,600 Hz to 2,300–2,400 Hz between pretest and posttest sessions in all contexts (task types, following vowel backness, and following vowel height). The FFI + CF group also outperformed the control group on familiar items in the picture-description task. Taken together, we interpret these results as a significant improvement in their pronunciation of English /ɹ/, which suggests that FFI in conjunction with CF might be an effective and efficient way to promote not only L2 grammar but also L2 pronunciation development, at both a controlled- and spontaneous-speech level, irrespective of vowel contexts. Previous pronunciation teaching studies drew, for the most part, on decontextualized instructional methods and showed an impact on students'

performance only at a controlled-speech level. By comparison, the findings of the current study reveal the benefits of communicatively oriented pronunciation teaching. In order for instruction to impact their communicative competence, it might be the case that learners need to process form jointly with meaning and practice it repetitively in communicatively authentic contexts (Celce-Murcia et al., 1996; Pennington, 1996; Segalowitz, 2003; Trofimovich & Gatbonton, 2006).

In contrast, the results revealed that the FFI-only group did not demonstrate any significant F3 decline in any context. Although the FFI activities were designed to promote learners' noticing and awareness of the target pronunciation form (through structured input, typographically enhanced input, and focused tasks), they proved insufficient on their own, without CF, to significantly impact the learners' developing L2 system. These results bring to the fore a crucial theoretical issue in SLA—the role of both positive and negative evidence in L2 pronunciation development; namely, L2 pronunciation development might require not only enhanced positive evidence (i.e., proactive FFI) but also immediate negative evidence from the teachers (i.e., reactive FFI). Even after they begin to gain awareness of the target pronunciation feature of English /ɹ/ as well as intensive exposure to the teachers' model pronunciation, learners might still need negative evidence in the form of CF from their teachers in order to (a) double-check whether their own pronunciation of English /ɹ/ is intelligible enough and (b) revise their own output in response to the teachers' model pronunciation. In conjunction with several research findings that Japanese learners of English gradually improve their production and perception abilities of English /ɹ/ over substantial length of residence in English-speaking countries (Larson-Hall, 2006), it is possible that FFI that provides both enhanced positive evidence and immediate negative evidence in an effective and efficient manner can expedite the rate of L2 pronunciation development (Doughty, 2003).

Although previous intensive perception training studies have produced evidence that learners' improvement in their use of familiar tokens can be transferred to novel tokens (e.g., Bradlow et al., 1997; Hardison, 2003; Iverson et al., 2005; Lively et al., 1993), the current study showed no group differences in the generalizability test in which the learners were asked to read four new non-minimally-paired words containing the English /ɹ/ in word-initial positions. Despite the lack of significance, both the FFI + CF group and the FFI-only group, however, noted small-to-medium effect sizes in comparison with the control group, suggesting that participants who received FFI treatments (i.e., ± CF) might have begun to apply their improved abilities to new contexts (to

words that did not appear during the FFI), but with considerable individual variance. This topic needs future research that will include more speech samples and more free-constructed measures to elicit L2 speech production at a spontaneous-speech level.

Conclusion

The current study took a first step toward investigating the value of FFI for supporting L2 pronunciation development. It did so via the integration of FFI, a research framework originally developed for L2 grammar studies, with interdisciplinary methodologies from L2 pronunciation research and phonology research. The study provided a number of noteworthy findings. First, a communicative focus on phonological form can benefit L2 pronunciation development. This is the case even with English /ɹ/, supposedly the most difficult sound for adult Japanese learners of English. Second, the impact of FFI on learners' interlanguage development was apparent not only at a controlled-speech level but also at a spontaneous-speech level, suggesting that FFI can promote not only development of a new metalinguistic representation of English /ɹ/ but also its internalization in a learner's L2 developing system. Third, although it is important to develop learners' selective attention toward the target pronunciation feature of English /ɹ/ through enhanced positive evidence in instructional input (i.e., proactive FFI), the learners still need immediate negative evidence (i.e., CF) in order to ascertain whether their output is perceived as sufficiently intelligible. The relative importance of CF may be due to the fact that, with respect to L2 pronunciation development, it is difficult for L2 learners to make online judgments on their own about the extent to which their interlanguage form is good enough (i.e., intelligible pronunciation).

In addition, the current study adopted unique methodological features with respect to outcome measures. Instead of adopting only human rating methods, the current study adopted both human rating methods and acoustic analyses. This technique enhanced the validity of our study by enabling us to (a) to include a sufficient number of participants and test materials to measure students' improvement at various levels and (b) to track with relative precision any changes over time in the speech properties of the target features. Therefore, we believe we have offered a model for a methodology that can foster reliable outcome measures in future L2 pronunciation classroom-based studies of an interventional kind, which should first (a) ask NS listeners to rate a small subset of speech tokens from the original data pool, then (b) carefully find

which speech properties relate to their rating scores, and, finally, (c) conduct acoustic analyses on the entire set of speech tokens to measure the effects of instruction on these relevant speech properties at a fine-grained level.

Limitations and Pedagogical Implications

Because the instructors were encouraged to give priority to developing their students' clear understanding of meaning (i.e., English argumentative skills) rather than form (i.e., English /ɹ/), the instructional treatment in the current study could not be as consistent as that in laboratory-based studies. However, in light of the our research goal—the *pedagogical* effectiveness of FFI on L2 pronunciation development— it seems imperative for attaining ecological validity to conduct research in classroom contexts, even though many variables cannot be controlled as much as in laboratory settings. Second, use of the timed picture-description task as the only means of eliciting learners' spontaneous-speech production in the current study is also a limitation. We call for more studies to refine and develop ecologically valid outcome measures specific to the context of pronunciation teaching studies. Third, the design of the generalizability test needs to be refined; a wide range of tasks should be included in the future to more appropriately investigate how learners can generalize newly acquired knowledge to novel contexts. Fourth, as perception training studies have generated long-term positive results, it would be important for future research to investigate the sustainability of FFI effectiveness over a longer period of time (see Derwing & Munro, 2005).

The current study is the first to show that teaching pronunciation forms embedded in meaningful contexts can enhance learners' L2 speech performance not only during controlled but also during more spontaneous production. By contrast, the effects of more explicit interventions that have been investigated in the past, such as explicit instruction on oral gestures (e.g., Elliott, 1997; Macdonald et al., 1994), remain unclear. Second, the role of teachers' immediate feedback might be relatively important for pronunciation teaching, because students need to (a) receive the teachers' feedback on the intelligibility of their output (negative evidence) and (b) practice the correct form in response to their teachers' model pronunciation (positive evidence). In this respect, we suggest that pronunciation recasts might be especially effective for L2 pronunciation development, but we recommend further research that compares a variety of CF types. Given that our findings were based on the specific case of L2 speech acquisition of English /ɹ/ by Japanese learners, we call for future research to replicate and extend the current research framework; in particular, it would be intriguing if future experimental studies investigated the effects of FFI

on L2 phonological development but with respect to other less salient sound contrasts such as the /æ/-/ɛ/ and /i/-/ɪ/ distinctions (Derwing & Munro, 2005) or suprasegmentals such as word stress (Field, 2005) and speech rate (Munro & Derwing, 2001).

<div align="right">Revised version accepted 31 October 2010</div>

Notes

1 This line of research also showed that F2 and transition duration can be secondary phonetic cues (for details, see Hattori & Iverson, 2009).

2 The posttest sessions in the current study could be considered as "short-delayed posttests" rather than "immediate posttests" according to the FFI research standards in L2 grammar studies (e.g., Mackey & Goo, 2007; Spada & Tomita, 2010). To our knowledge, none of the previous pronunciation teaching studies adopted delayed posttesting measures (see Derwing & Munro, 2005).

3 In addition, to avoid any lexical familiarity effect (i.e., L2 speech production is relatively better when learners are familiar with target words; see Flege, Frieda, Walley, & Randazza, 1998), all participants were asked if any words in the test materials were unfamiliar to them. All of the target words were found to be quite familiar to the learners.

4 Given that learners' performance of minimally paired words including English /r/ in word-medial positions was not the focus of the current study, the decision was made to exclude two minimally paired words (i.e., "arrive" and "correct") from the original analysis.

5 Lively et al. (1993) actually showed that their tailored perceptual training, which focused on English /ɹ/ and /l/ occurring only in its most difficult position in words (prevocalic), improved the perception abilities of Japanese learners of English and confirmed their transferred effects to other relatively easy contexts such as English /ɹ/ and /l/ on postvocalic positions.

6 Although conducting listening sessions in this way does not correspond to a real-life situation in which listeners have only one opportunity to listen and understand interlocutors, note that it would have been otherwise very difficult for listeners to rate only the quality of English /ɹ/ rather than basing their judgment on the whole word. In fact, all of the raters reported that the talkers tended to mispronounce not only English /ɹ/ but also many other segmentals such as /æ/ in *ram*.

7 According to Cohen (1988), effect sizes are roughly classified as small ($0.20 \leq d < 0.50$), medium ($0.50 \leq d < 0.80$), or large ($0.80 \leq d$). In all cases, control group means were used to calculate between-group contrasts and pretest/posttest scores were used to calculate within-group contrasts.

References

Aoyama, K., Flege, J., Guion, S., Akahane-Yamada, & Yamada. (2004). Perceived phonetic dissimilarity and L2 speech learning: the case of Japanese /ɹ/ and English /l/ and /ɹ/. *Journal of Phonetics, 23*, 233–250.

Best, C., & Strange, W. (1992). Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of Phonetics, 20*, 305–330.

Boersma, P., & Weenik, D. (2009). *Praat: Doing phonetics by computer*. Retrieved September 2009 from http://www.praat.org

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

Bradlow, A., Akahane-Yamada, R., Pisoni, D., & Tohkura, Y. (1999). Teaching 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., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*, 707–729.

Bradlow, A., Pisoni, D., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /ɹ/ and /l/. *Journal of the Acoustical Society of America, 101*, 2299–2310.

Carpenter, H., Jeon, K. S., MacGregor, D., & Mackey, A. (2006). Learners' interpretations of recasts. *Studies in Second Language Acquisition, 28*, 209–236.

Celce-Murcia, M., Brinton, D., & Goodwin, J. (1996). *Teaching pronunciation: A reference for learners of English to speakers of other languages*. Cambridge: Cambridge University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Couper, G. (2006). The short and long-term effects of pronunciation instruction. *Prospect, 21*, 46–66.

DeKeyser, R. (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.

Derwing, T., & Munro, M. (2005). Second language accent and pronunciation teaching: A research–based approach. *TESOL Quarterly, 39*, 379–397.

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

Derwing, T., & Munro, M., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning, 48*, 393–410.

Doughty, C. (2003). Instructed SLA: Constraints, compensation, and enhancement. In M. Long, & C. Doughty (Eds.), *Handbook of second language acquisition* (pp. 257–310). Malden, MA: Blackwell.

Doughty, C., & Williams, J. (Eds.). (1998). *Focus on forms in classroom second language acquisition*. Cambridge: Cambridge University Press.

Ellis, R. (2001). Investigating form-focused instruction. *Language Learning, 51*, 1–46.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition, 27*, 141–172.

Ellis, R. (2006). Researching the effects of form-focussed instruction on L2 acquisition. *AILA Review, 19*, 18–41.

Ellis, R., Basturkmen, H., & Loewen, S. (2001). Learner uptake in communicative ESL lessons. *Language Learning, 51*, 281–318.

Ellis, R., & Sheen, Y. (2006). Re-examining the role of recasts in L2 acquisition. *Studies in Second Language Acquisition, 28*, 575–600.

Elliott, A. (1997). On the teaching and acquisition of pronunciation within a communicative approach. *Hispania, 80*, 95–108.

Espy-Wilson, C. (1992). Acoustic measures for linguistic features distinguishing the semivowels /wjrl/ in American English. *Journal of Acoustical Society of America, 92*, 736–751.

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly, 39*, 399–423.

Flege, J. (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.

Flege, J., Frieda, E., Walley, A., & Randazza, L. (1998). Lexical factors and segmental accuracy in second-language speech production. *Studies in Second Language Acquisition, 20*, 155–188.

Flege, J., Munro, M., & MacKay, I. R. A. (1995). Effects of age of second-language learning on the production of English consonants. *Speech Communication, 16*, 1–26.

Flege, J., Takagi, N., & Mann, V. (1995). Japanese adults learn to produce English /ɹ/ and /l/ accurately. *Language and Speech, 38*, 25–55.

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

Han, Z., Park, E., & Combs, C. (2008). Textual enhancement of input: Issues and possibility. *Applied Linguistics, 29*, 597–618.

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

Hardison, D. (2010). Visual and auditory input in second-language speech processing. *Language Teaching, 43*, 84–95.

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

Ingram, J. C., & Park, S. (1998). Language, context, and speaker effects in the identification and discrimination of English /ɹ/ and /l/ by Japanese and Korean listeners. *Journal of the Acoustical Society of America, 103*, 1161–1174.

Ioup, G., Boustagi, E., El Tigi, M., & Moselle, M. (1994). Reexamining the critical period hypothesis: A case study of successful adult SLA in a naturalistic environment. *Studies in Second Language Acquisition, 16*, 73–98.

Iverson, P., & Kuhl, P. (1996). Influences of phonetic identification and category goodness on American listeners' perception of /ɹ/ and /l/. *Journal of the Acoustical Society of America, 92*, 1130–1140.

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., Kuhl, P., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, et al. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*, 847–857.

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford: Oxford University Press.

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics, 23*, 83–103.

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

Ladefoged, P. (2003). *Phonetic data analysis: An introduction to instrumental phonetic fieldwork*. Oxford: Blackwell.

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. (2006). What does more time buy you? Another look at the effects of long-term residence on production accuracy of English /ɹ/ and /l/ by Japanese speakers. *Language and Speech, 49*, 521–548.

Levis, J. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly, 39*, 367–377.

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

Logan, J., Lively, S., & Pisoni, D. (1992). Training Japanese listeners to identify English /ɹ/ and /l/: A first report. *Journal of the Acoustical Society of America, 89*, 874–886.

Lyster, R. (1998a). Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classrooms. *Language Learning, 48*, 183–218.

Lyster, R. (1998b). 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: Benjamins.

Lyster, R., & Saito, K. (2010a). Corrective feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition, 32,* 265–302.

Lyster, R., & Saito, K. (2010b). Interactional feedback as instructional input: A synthesis of classroom SLA research. *Language, Interaction and Acquisition, 1,* 276–297.

Macdonald, D., Yule, G., & Powers, M. (1994). Attempts to improve English L2 pronunciation: The variable effects of different types of instruction. *Language Learning, 44*, 75–100.

Mackey, A., Gass, S., & McDonough, K. (2000). How do learners perceive interactional feedback? *Studies in Second Language Acquisition, 22*, 471–497.

Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 407–452). Oxford: Oxford University Press.

Mochizuki, M. (1981). The identification of /l/ and /ɹ/ in natural and synthesized speech. *Journal of Phonetics, 9*, 283–303.

Moyer, A. (1999). Ultimate attainment in L2 phonology: The critical factors of age motivation and instruction. *Studies in Second Language Acquisition, 21*, 81–108.

Munro, M., & Derwing, T. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition, 23*, 451–468.

Munro, M., & Derwing, T. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System, 34*, 520–531.

Neri, A., Mich, O., Gerosa, M., & Giuliani, D. (2008). The effectiveness of computer assisted pronunciation training for foreign language learning by children. *Computer Assisted Language Learning, 21*, 393–408.

Nicholas, H., Lightbown, P., & Spada, N. (2001). Recasts as feedback to language learners. *Language Learning, 51*, 719–758.

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

Pennington, M. (1996). *Phonology in English language teaching*. London: Longman.

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

Polka, L., & Strange, W. (1985). Perceptual equivalence of acoustic cues that differentiate /ɹ/ and /l/. *Journal of the Acoustical Society of America, 78*, 1187–1197.

Rivers, W., & Temperley, M. (1978). *A practical guide to the teaching of English as a second language*. Oxford: Oxford University Press.

Robinson, P. (1996). Learning simple and complex rules under implicit, incidental rule-search conditions, and instructed conditions. *Studies in Second Language Acquisition, 18*, 27–67.

Saito, K. (2011a). Re-examining the role of explicit phonetic instruction in native-like and comprehensible pronunciation development: An instructed SLA approach to L2 phonology. *Language Awareness, 20*(1).

Saito, K. (2011b). Towards identifying teaching and learning priorities for acquiring intelligible pronunciation. Manuscript submitted for publication.

Segalowitz, N. (2003). Automaticity and second languages. In C. Doughty & M. Long (Eds.), *The handbook of second language acquisition* (pp. 382–408). Oxford: Blackwell.

Setter, J., & Jenkins, J. (2005). Pronunciation. *Language Teaching, 38*, 1–17.

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 /ɹ/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics, 3*, 243–261.

Spada, N. (1997). Form-focused instruction and second language acquisition: A review of classroom and laboratory research. *Language Teaching, 29*, 73–87.

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning, 60*, 263–308.

Trofimovich, P., & Gatbonton, E. (2006). Repetition and focus on form in L2 Spanish word processing: Implications for pronunciation instruction. *Modern Language Journal, 90*, 519–535.

VanPatten, B. (2004). *Input processing and grammar instruction in second language acquisition*. Norwood, NJ: Ablex.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. FFI Treatment.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.