CORE
Metadata, citation and similar papers at core.ac.uk
Provided by University of Richmond

Masthead Logo

University of Richmond
UR Scholarship Repository

Latin American, Latino and Iberian Studies Faculty Publications

Latin American, Latino and Iberian Studies

2018

Pronunciation Instruction Can Improve L2 Learners' Bottom-Up Processing for Listening

Elizabeth M. Kissling
University of Richmond, ekisslin@richmond.edu

Follow this and additional works at: https://scholarship.richmond.edu/lalis-faculty-publications

Part of the Latin American Languages and Societies Commons, and the Linguistics Commons

This is a pre-publication author manuscript of the final, published article.

Recommended Citation

Kissling, Elizabeth M., "Pronunciation Instruction Can Improve L2 Learners' Bottom-Up Processing for Listening" (2018). *Latin American, Latino and Iberian Studies Faculty Publications*. 150.
https://scholarship.richmond.edu/lalis-faculty-publications/150

This Post-print Article is brought to you for free and open access by the Latin American, Latino and Iberian Studies at UR Scholarship Repository. It has been accepted for inclusion in Latin American, Latino and Iberian Studies Faculty Publications by an authorized administrator of UR Scholarship Repository. For more information, please contact scholarshiprepository@richmond.edu.

Pronunciation Instruction Can Improve L2 Learners' Bottom-Up Processing for Listening

ABSTRACT

Listening is widely regarded as an important skill that is difficult and necessary to teach in L2 classrooms. Listening requires both top-down and bottom-up processing, yet pedagogical techniques for the latter are often lacking. This study explores the efficacy of pronunciation instruction (PI) for improving learners' bottom-up processing. The study recruited 116 relatively novice learners of Spanish as a foreign language and provided the experimental groups with brief lessons in PI emphasizing segmental or suprasegmental features followed by production-focused or perception-focused practice. Learners' bottom-up processing skill was assessed with a sentence-level dictation task. Learners given PI on suprasegmental features followed by perception-focused practice found target language speech to be more intelligible than controls, indicating that they had improved their bottom-up processing. However, learners given PI on segmental features followed by production-focused practice found target language speech to be more comprehensible. The results indicate that PI is a worthwhile intervention for reasons that go beyond pronunciation, even when instructional time is limited, and that a range of features and practice types should be included in PI to improve listening skills.

Keywords: listening; pronunciation; pedagogy; classroom-based research; phonetics/phonology; Spanish

LISTENING IN A SECOND LANGUAGE

Listening to fluent speech is central both to communicating successfully in a second language (L2) and also acquiring the language through exposure to rich input. Though necessary for communication and learning, listening is also a complex and highly challenging skill that requires much of leaners (Brown, 2013; Cutler, 2001; Graham, 2006; Goh, 2000; Vandergrift & Goh, 2012). First, successful listeners must call upon prior knowledge about the topic and context in order to make useful predictions about what they might hear. That is, they must activate prior knowledge that is likely to be relevant for the particular listening task at hand in order to understand the overall message without necessarily understanding every word. This is known as 'top-down' processing (Field, 2004). Learners must also simultaneously engage in 'bottom-up' processing, calling upon their knowledge about the L2 linguistic system in order to segment speech, identify words, and parse what they hear. Bottom-up processing allows listeners to build up an understanding of the intended message directly from the incoming speech stream (Field, 2004). However, most L2 learners are not adept at bottom-up processing; they have difficulty segmenting the speech stream into meaningful units, keeping enough in their working memory to piece together related units, identifying words they know, and disambiguating homophones based on immediate context, amongst other things (Brown, 2013; Cutler, 2001; Goh, 2000). Their processing difficulties are multifaceted. For one, L2 listeners tend to activate 'phantom' words (Broersma & Cutler, 2008) while engaged in online processing. That is, they consider multiple word candidates that are phonologically plausible but do not actually exist in the L2, which complicates the word recognition process. They also allow top-down expectations to overrule bottom-up information detected, tending to substitute a word they know for the word actually present if it is unknown to them, which Field (2004) calls a 'lexical effect.' In general, L2 learners are more reluctant than L1 speakers to change their predictions when confronted with conflicting evidence in the signal. In other words, they hear what they expect to hear. In an extensive review of the L2 listening research, Cutler (2001) concluded that listeners exploit different heuristic procedures when segmenting speech. Those L2 processing procedures, initially dictated by the structure of a learner's L1, are applied automatically to the L2 even though they are often inefficient.

Since successful listening requires listeners to call upon a wealth of knowledge quite rapidly, it is not surprising that learners find listening to be difficult. Interestingly, they tend to ascribe their difficulty to bottom-up processing problems rather than top-down. For example, the L1 English-L2 French learners (n = 595) in Graham's (2006) survey reported that two of their main problems were dealing with fast speech rate and identifying individual words in the speech stream, both of which are bottom-up processing issues. Similarly, the bottom-up processes of recognizing known words and breaking the speech stream into chunks were identified as stumbling blocks in Goh's (2000) survey of adult L1 Chinese-L2 English learners, particularly for low-ability listeners. Vandergrift and Baker (2015) conducted an exploratory path analysis with L1 English-L2 French adolescents (n = 157) to investigate the relative importance of a number of variables that predict L2 listening ability. Their results suggested that auditory discrimination, a bottom-up processing skill, was more important than other variables in early stages of acquisition. Their results echoed those of Wilson et al. (2011), who found a moderate correlation between auditory discrimination and L2 listening comprehension, which led Vandergrift and Baker (2015) to argue that "some attention to consciousness-raising at the sound segment level would be useful, particularly with learners at lower levels of language proficiency" (p. 411). Precisely how best to instruct learners to improve their bottom-up processing, however, is an open question.

INSTRUCTION FOR BOTTOM-UP PROCESSING

Listening is largely a learner-internal process, making it opaque to instructors and particularly challenging to teach. Teachers often provide opportunities for learners to strengthen their listening skills by engaging in practice, but not all teachers explicitly instruct learners on precisely how they might improve in listening (Brown, 2013, p. 36). Novice learners in particular need to improve their bottom-up processing and could benefit from explicit instruction, as well as practice, for developing those skills (Field, 2003). Research-based teaching guides tend to emphasize top-down metacognitive listening strategies (e.g. Vandergrift & Goh, 2012) rather than bottom-up instruction. A number of studies have suggested that such top-down strategies improve listening comprehension. For instance, Yeldham (2016) reported that teaching strategies led to more improvement in listening comprehension than an approach that balanced top-down and bottom-up skills, even though the balanced approach led to better bottom-up processing in particular tasks. However, Yeldham's (2016) study recruited intermediate EFL learners, and the same results might not have been found with less proficient learners, who tend to rely more on bottom-up processing (Goh, 2000; Vandergrift & Baker, 2015). Siegel and Siegel's (2015) study with lower-intermediate EFL university learners ($n = 43$) suggested that bottom-up skills activities like highlighting connected speech in transcripts and counting words improved their listening. Unfortunately, such bottom-up skills work is not often incorporated into listening instruction (Siegel & Siegel, 2015). One potential way for teachers to incorporate bottom-up processing would be via pronunciation instruction (PI). Increasing learners' knowledge about the sound system of the L2 might help them segment the speech stream and identify words. To date, scant research has investigated the effect of PI on bottom-up processing to support L2 listening, but there is reason to believe it could be beneficial.

Direct evidence that learners can be taught to better perceive speech in their L2 comes from perceptual training experiments. They provide learners with intensive exposure to a limited number of target sounds (see Logan & Pruit, 1995 for methods) and then test their ability to identify or discriminate the target sounds in contexts identical or similar to those trained. For instance, a wealth of studies (Sakai & Moorman, 2018; Thomson, 2012) have trained Japanese speakers to perceive the English /ɹ/ - /l/ contrast in order to differentiate minimal pairs like right / light (e.g. Bradlow, Pisoni, Akahane-Yamada & Tohkura, 1997). The target sounds most often selected for these training experiments are vocalic and consonantal segments (Shin & Iverson, 2013), either in simple phonemic contrasts or as a class of sounds, such as the full vowel system or a subset of difficult vowels (e.g. Iverson & Evans, 2009; Nishi & Kewley-Port, 2007, 2008). For instance, Thomson (2012) trained 26 Mandarin speakers on 10 English vowels, presenting them with 200 tokens of the vowels spoken by 20 native speakers in the context of [b_] over the course of eight short training sessions. Participants' perception of the vowels significantly improved after training. Fewer studies have attempted perceptual training on targets other than segments, though there are some exceptions such as Shin and Iverson's (2013) study on prosody, which trained English learners to identify sentence focus based on stress.

During training leaners are asked to identify and/or discriminate difficult L2 sounds. Discrimination training is thought to target the auditory-phonetic level, helping learners attune their auditory-phonetic perceptual processing to the L2, whereas identification training is thought make learners categorize the sounds they hear and thus help them improve their mental representations of the sounds at the phonological level (see Shinohara & Iverson, 2018, for a recent study comparing training types). Under either training condition, a putatively critical element of the training is that the stimuli be phonetically variable like natural speech, i.e.

multiple tokens produced by multiple talkers. This variability is thought to boost learning and produce effects that generalize to new stimuli and new contexts. The High Variability Phonetic Training (HVPT) paradigm, explained in the seminal studies of Logan, Lively, and Pisoni (1991) and Lively, Logan, and Pisoni (1993), has become the gold standard for perceptual training. HVPT has been found to support learning of difficult target language contrasts more than exposure alone, even for very experienced L2 users with intensive naturalistic input (Iverson, Pinet & Evans, 2012; Shinohara & Iverson, 2018), perhaps because of the focused attention it requires (Logan et al., 1991).

For all its benefits, HVPT is not a panacea for learners' difficulties with L2 sounds. Some have argued that HVPT helps learners get faster at applying their existing knowledge while processing real speech but does not actually change their lower-level processing routines (Iverson, Hazan & Bannister, 2005). Others have noted that the perceptual benefits incurred from HVPT do not transfer to new phonetic contexts in all cases (e.g. Thomson, 2012). Finally, while intensive, isolated training and testing can improve perception, such training conditions are not like most communicative classroom environments where learners ideally are engaged in authentic discourse during meaning-making activities. Even the 'second wave' of training studies—experiments that have moved beyond simple phonemic contrasts, incorporated sophisticated computer programming, and designed training to be more learner-centered (Thomson, 2012)—rarely presents stimuli in communicative contexts, which puts HVPT at odds with current teaching methodologies. The same can be said of pronunciation instruction more generally (Derwing & Munro, 2005). Indeed, it is precisely the element of drawing learners' attention *from* meaning *to* form that seems to confer a benefit in sound perception. One promising avenue for perceptual training is a recent, perhaps counterintuitive, finding that presenting stimuli in increasingly larger linguistic contexts actually makes the training both easier and more beneficial (Kewley-Port et al., 2009).

Many of the perceptual training experiments undertaken in the 1990s were inspired by the work of James Flege and his Speech Learning Model (SLM) (Sakai & Moorman, 2018), one of the most widely-used models of L2 phonological acquisition. The SLM claims that accurate perception of L2 sounds is necessary for accurate production of the same sounds (Flege, 1995). Thus, perceptual training was explored as a means of facilitating formation of accurate phonological representations in the mind and thereby facilitating accuracy in pronunciation (Sakai & Moorman, 2018). In addition to the SLM, other accounts of L2 phonological acquisition also posit a strong relationship between perception and production (Akerberg, 2005; Broselow & Park, 1995; Colantoni & Steele, 2008), though in practice they are often investigated separately (Leather & James, 1996; Leather, 1999). A meta-analysis of 111 perceptual training studies published between 1988 and 2013 that measured changes in production concluded that perception training has led to small production gains, which tend to be greater when learners are at lower proficiency levels and the training is accompanied by phonetic instruction (Sakai & Moorman, 2018, p. 212).

Perceptual training of some sort is often included as part of classroom-based pronunciation instruction (PI) as well. Though much of the extant research on PI unfortunately underreports precise details about pedagogical practices (Lee, Jang & Plonsky, 2015; Thomson & Derwing, 2015), it seems that most PI involves some perceptual component. Some studies employ the HVPT, some provide discrimination and/or identification practice of target phones, and some use speech software to display visual feedback (e.g. spectrograms and waveforms) delivered simultaneously with perceptual practice (Lee, Jang & Plonsky, 2015; Thomson &

Derwing, 2015). Still, the outcome most of interest in PI studies is usually learners' pronunciation, not their perception per se.

Though the field "should be primarily concerned with *helping learners become more understandable*," much PI research emphasizes accent and privileges native-likeness in production (Thomson & Derwing, 2015, p. 327). Studies have shown that PI can improve learners' L2 pronunciation in terms of reduced accent, assessed with both global accent ratings and measures of fine phonetic details (Saito, 2012). PI has also been shown to increase the intelligibility and comprehensibility of L2 speech (Saito, 2012). Lee, Jang and Plonsky's (2015) meta-analysis (total n = 2,782) found medium to large effect sizes of PI in classroom studies, though the authors caution that the effectiveness of PI for improving pronunciation may be overstated for several reasons: sample sizes have been small, larger effect sizes have been found in studies that employed more controlled tasks, smaller effect sizes have been found in studies that recruited real control groups, and there has been a bias towards publishing statistically significant results. Indeed, some studies with larger samples and control groups have reported little effect of PI on pronunciation compared with other types of instruction (e.g. Kissling, 2013). Still, recent studies are continuing to report positive effects of PI on pronunciation (e.g. Huensch, 2016), even when the PI is quite brief (e.g. Gordon & Darcy, 2016). It is also reasonable to postulate that if PI can help learners improve their speech production, it can also help their speech perception. Perception and production have been characterized as mutually facilitative (e.g. Leather, 1999), meaning that improvement in either skill can spark improvement in the other. If PI improves learners' pronunciation, the Speech Learning Model would predict that it has also improved their perception (Flege, 1995). Such perceptual gains could, in Cutler's (2001) terms, help learners develop more efficient heuristics for bottom-up processing of target language speech. This may be especially true for novice learners, who typically exhibit more rapid improvement in pronunciation after PI than more experienced learners (Derwing & Munro, 2005) and may experience the greatest benefit from perceptual training (Sakai & Moorman, 2018).

PI has been shown to improve the perception outcomes of learners of Spanish as a foreign language (SFL), who constitute the primary experimental group in this study. Ausín and Sutton (2010) found that advanced SFL learners (*n* = 39) were better able to detect English-accented pronunciations of some consonantal target phones in Spanish after taking a semester-long course on pronunciation and phonetics. Kilpatrick and McLain Pierce (2014) found that much more limited PI also improved perception. The advanced SFL learners (*n* = 17) in their study were better able to correctly perceive diphthongs as single syllable nuclei rather than two syllables after just ten minutes of explicit instruction. Kissling (2015) provided novice and intermediate SFL learners (*n* = 46) two hours of phonetics-focused PI and found that instructed learners garnered a small advantage in detecting English-accented pronunciations of some consonantal target phones in Spanish, even weeks after receiving the instruction. Rasmussen and Zampini (2010) expanded studies of perception after PI to the realm of L2 dialectal variation and to sentence-level tasks as opposed to the word-level stimuli of the other studies mentioned here. They provided third and fourth year SFL learners (*n* = 16) with six weeks of PI focused on specific features of Andalusian Spanish and then assessed the intelligibility and comprehensibility of that dialectal variety for students. Intelligibility refers to what is actually heard (as measured by a dictation task), whereas comprehensibility refers to listeners' perceived difficulty in comprehending speech (as measured by their ratings). Compared to uninstructed controls, the learners who received PI found some Andalusian speech features to be more intelligible and comprehensible. In sum, PI has been shown to help relatively advanced SFL leaners better perceive particular features targeted in the PI. More research is necessary to determine whether PI can help more novice learners improve listening more generally.

RESEARCH QUESTIONS

So far it has been argued that listening is important, bottom-up processing is essential for listening but difficult for beginning L2 learners, and PI might be useful because it attunes learners to sounds of the L2. Levis (2016) has urged PI researchers to expand their focus from the question of what (what to teach or if to teach) to questions of why, how and for whom PI might best work (p. 6). This study addresses who might benefit from PI by recruiting beginning learners. This expands the extant research because learners recruited in PI studies are more often intermediate to advanced, especially in SFL studies. The current study's main research questions, though, relate to the why and how of PI.

RQ 1 (Why): Does PI make target language speech more intelligible or comprehensible for L2 learners?

PI has been shown to confer an advantage in L2 pronunciation and in perception tasks that typically present a limited number of target sounds in relative isolation. The hypothesis of the current study is that since PI attunes listeners to the target language sound system, it can also strengthen their bottom-up processing skills, in particular speech segmentation and word identification, and help them find target language speech more intelligible and comprehensible.

RQ 2 (How): What type of practice most impacts the intelligibility and comprehensibility of target language speech for L2 listeners: perception-focused or production-focused practice?

Field (2003) highlighted the utility of dictation practice for teaching the bottom-up skills that support lexical segmentation in L2 speech. Kissling (2013) found that the input, practice, and feedback typically provided during PI were equally as important as explicit phonetics instruction for improving learners' L2 pronunciation. It is hypothesized that practice will be equally important for listening, and that practice will be more effective if it requires learners to apply PI in the perceptual domain by engaging in dictation tasks, as Field (2003) suggests.

RQ3 (What): Which target features of PI most impact intelligibility and comprehensibility of target language speech for L2 listeners: segmental or suprasegmental features?

So far "very few empirical investigations have addressed the relative effectiveness of PI on these two feature types" (Lee, Jang & Plonsky, 2015, p. 348), and this study seeks to address that gap. Much research on PI, particularly in SFL, has studied segmental targets. Segmental targets are isolated vocalic and consonantal phonemes. Consonants especially have been investigated extensively in the SFL context, perhaps because precision with consonants is thought to correlate with accent (see Kissling, 2013, for a review), and also perhaps because they are easier to teach and measure (Levis, 2005). However, some studies (e.g. Hahn, 2004) suggest that suprasegmental features might impact pronunciation more than segmentals. Suprasegmental features are those that spread across multiple phonemes, for example word stress, rhythm, and resyllabification across words ('linking'). For instance, Gordon and Darcy's (2016) study found that explicit PI focused on suprasegmental features (stress, rhythm, reductions, and linking) was more effective than explicit PI on segmental features (vowels) and implicit instruction for increasing the comprehensibility of ESL learners' ($n = 12$) speech. Following these studies, the hypothesis posited here is that suprasegmental PI will improve learners' speech segmentation and word identification more than segmental PI and so will increase the intelligibility of the target language speech they hear. Given the paucity of research on comprehensibility, no hypothesis is made about the relative effect of each type of PI on comprehensibility.

In sum, the hypotheses are that segmental and suprasegmental PI will make target language speech more intelligible and comprehensible for novice SFL listeners, especially when the instruction includes listening practice, though the effect size is likely to be driven by the instructional time available.

METHOD

Context and Participants

This quasi-experimental study was carried out in a small, private university in the Mid-Atlantic United States. During two semesters, all the students enrolled in an accelerated beginning Spanish FL course were invited to participate. All agreed to participate, but 6 were removed from the data set due to not being present during the posttest, resulting in a total of 116 participants (39 F, 77 M). All were 18 – 21 years of age and L1 speakers of English. The course was designed for students who had some previous Spanish instruction (typically 1 – 2 years) but lacked the requisite proficiency to enroll in an intermediate course, as determined by a placement test. None were absolute beginners, but none were rated as intermediate by the placement test. During the first week of class, instructors re-assigned any misplaced students, basing their judgments on classroom interactions and written work. Impressionistically, instructors rated the students as novice-mid to novice-high proficiency on the ACTFL scale, but as with any intact class study, individuals' true proficiency was varied.

Four intact classes were assigned to four experimental conditions. Some received segmental PI while others received suprasegmental PI, and half of each instructional group received a particular kind of practice. That is, one group (n = 12) received segmental PI followed by production-focused (pronunciation) practice. Another group (*n* = 14) received segmental PI followed by perception-focused (listening) practice. A third group (*n* = 13) received suprasegmental PI followed by production-focused practice. The final group (*n* = 11) received suprasegmental PI followed by perception-focused practice.

An L2 speaker of Castilian Spanish with expertise in Spanish phonetics was the instructor for all four classes. The first two groups were taught in one year and the second two groups were taught the following year. The curriculum and daily teaching methods were identical across those two years, with the exception of the PI of interest here. The control group (*n* = 66) was made up of five intact classes recruited during both years and taught by three instructors. One was an L1 speaker of Argentinian Spanish and two were L2 speakers of Castilian Spanish.

Instruction

The students met with instructors three times per week for 45 minutes to engage in a variety of communicative activities. They also met twice weekly with advanced students for oral practice of grammar and vocabulary. The PI was delivered in class in four 20-minute modules, during weeks 4, 6, 8 and 10 across the 14-week semester. The phonetic targets selected for instruction (see Appendix A) have been investigated in previous studies and were thought to be among the features that could impede word identification and speech segmentation. The suprasegmental instruction focused on linking, diphthongs, and synalepha. The segmental instruction focused on taps and approximants (see Appendix A), all segments known to be particularly problematic in L2 Spanish pronunciation (see Kissling, 2013 for a review) and perception (see Kissling, 2015 for a review).

An instructor-driven, low-tech approach was chosen because human-delivered PI instruction has been found to produce larger effects than technology-delivered instruction (Lee, Jang & Plonsky, 2015). The instructional procedures were as follows. The instructor read aloud the title of the module in Spanish, which named the targets without using unnecessary phonetics jargon (Appendix B). The instructor then slowly pronounced a word or phrase that exemplified the PI target and repeated the same phrase three times, with natural speed. Students repeated the instructor's pronunciation model, in chorus, several times. The same procedure was followed for 20 phrases that featured the PI target contextualized within familiar vocabulary. The phrases were represented pictorially and with a transcription in standard orthography. Additionally, suprasegmental features were denoted by underlining in the transcription, and segmental features were denoted by phonetic transcriptions of the target sounds (Appendix B). Modules 2 – 4 were begun with a one-minute review of several phrases included in the previous modules.

The last 10 minutes of each module were dedicated to practice. The production-focused practice groups repeated phrases containing the target features, following the instructor's modeling of target-like pronunciation. This production practice was done first in chorus, and then individually as each student was prompted, in turn, by the instructor. The instructor gave feedback on individuals' pronunciation in this way: target-like pronunciations were acknowledged with a head nod and smile. After non-target like pronunciations, the instructor modeled the phrase again twice, once slowly and once at a natural speed, then prompted the student to try again. If a student could not produce a target-like phrase after multiple attempts, the instructor replied with an encouraging "*casi*" 'almost.'

The perception-focused practice groups listened to target language speech that highlighted the target features while they filled in blanks on partial transcriptions (Appendix C) and then received feedback on their accuracy. The first audio recording was an interview included as part of the textbook materials. The others were idioms read by the instructor because the textbook materials were not suitable (Appendix C). The control groups received no such PI or targeted practice. Otherwise, the curricula were identical except for normal variations in instructor style.

With PI and any other intervention, “instructional costs (time and energy) must be weighed against their potential benefits for L2 learners” (Lee, Jang & Plonsky, 2015, p. 349). This study allotted just 1.3 hours to PI, less than the median length of instruction in previous PI studies (4.25 hours; Lee, Jang & Plonsky, 2015) but more than some studies that have found no effect of instruction (0.5 hours or less; see Saito, 2012) and some studies that have reported a positive effect of very short interventions focused on just one target feature (e.g. Kilpatrick & McLain Pierce, 2014). The present study sacrificed potential analytical power in an effort to accommodate to the real time constraints of the instructed environment, allotting perhaps the most minimal instructional time likely to make a difference for multiple features and a sentence-level task.

Assessment

The pre-test was a sentence-level dictation task given to all students in the second week of the semester. As explained by Siegel and Siegel (2015), dictation tasks are useful to assess bottom-up processing because “they require listeners to exercise and provide visual evidence of phoneme perception and parsing abilities that are not explicitly evident when other instruments such as tests of overall listening comprehension are used” (p. 647). An identical dictation task was administered in the last week of the semester (week 14) as a posttest. The audio presented controlled, fluent, sentence-level speech: recordings of nine sentence-length common idioms, the first two of which were for practice only (Appendix D). The idioms included 20 segmental target features and 20 suprasegmental target features, 10 of each category (reducing and linking, diphthongs and synalepha, taps, approximants) (Appendix D). These targets were not contained in words that had been previously practiced in the phonetics modules. Many of the words in the task were unfamiliar to students. The idioms were spoken by nine different L1 speakers of Spanish who lived in the US and spoke with Latin American dialects. None exhibited non-standard realizations of the target features. A variety of unfamiliar dialects were included so as to isolate the effect of PI from the potential effect of familiarity with a particular dialect. The speakers were recorded in a quiet room with a studio quality microphone and digitized into a wav format (44kHz, 16 bit quantization).

During the dictation task, students listened to the recordings and attempted to complete the partial transcription provided on a paper answer sheet (Appendix D). They were instructed to write down every word, syllable, and sound they perceived, regardless of whether they knew the meaning of the words. Students heard each idiom four times, with a 3-second pause between each repetition. Between idioms, the recording was paused until all students finished writing. On their answer sheet they also rated how difficult each idiom was to understand, on a scale of 1 (very difficult) to 7 (very easy).

Coding and Data Analysis

Comprehensibility was assessed with students' ratings of difficulty for each idiom (scale 1 – 7). Intelligibility was assessed with the number of words that students were able to correctly transcribe. Each word of the idioms was marked as either correctly or incorrectly transcribed. No points were deducted if students inserted words erroneously. Given that the students were relative novices, some leeway was made for words that appeared to be likely heard correctly but transcribed incorrectly, following Yeldham's suggestions for scoring dictation tasks (2017). For instance, *cebada* 'barley' was coded as correctly transcribed if a student wrote *sebada* or *sevada*, because both transcriptions indicate that the sounds of the word were correctly perceived though misspelled due to the student's lack of familiarity with the word or limited knowledge of orthographic conventions. On the other hand, *cebada* was coded as incorrectly transcribed if a student wrote *semana* or *cepada* because both transcriptions indicate that the sounds of the word may have been incorrectly perceived.

As the effect of instruction was expected to differ according to outcome measure (intelligibility or comprehensibility) and vary across focus of practice (production or perception) and instructed feature (segmental or suprasegmental), the groups' data were analyzed separately, in that order: outcome, practice, and instructed feature. To model the data, mixed-effects models (MEMs) were employed because of several advantages the method yields over traditional multiple regression or ANOVA methods: (1) by-subject and by-item analyses can be done simultaneously, so as to generalize across people and items within a single analysis; (2) accuracy for each individual word for each participant is included in the analysis rather than adding up across multiple words to obtain a single value for each participant; and (3) MEMs properly model the multilevel structure of the data (e.g. subject-level variables such as instructional group vs. variables such as test time, which varies both by subject and by item) and are therefore not subject to the assumption of independence of observations as are multiple regression or ANOVA (Baayen, Davidson & Bates, 2008; Linck & Cunnings, 2015).

The binomial logistic MEMs for intelligibility (word accuracy scored 0, 1) were conducted with the 'lme4' package version 1.1-12 (Bates, Maechler, Bolker & Walker, 2015), and the ordinal MEMs for comprehensibility (idiom Likert ratings ranged 1 – 7) with the 'ordinal' package version 2015.6-28 (Christensen, 2015) in R version 3.3.2 (R Core Team, 2016). One irregular response of '1.5' was removed from the ordinal dataset. Percentage estimates of effect sizes were obtained with the effects package version 3.1-2 (Fox, 2003), and reported effect sizes (percentage change in the probability of either an accurate response or a change in response on the Likert scale, respectively) for a particular variable were averaged across all other terms in the model. Logistic MEMs for word accuracy analyses were run using the 'bobyqa' optimizer.

For all MEMs, fixed effects included time (pretest, posttest; centered on pretest), and either pronunciation instruction (yes, no; centered on no) or instructional type (segmental, suprasegmental, none; centered on none). For the logistic MEMs of intelligibility (word accuracy of transcription), the covariate of word category (contains a phonetic feature of interest (FI), does not contain a phonetic feature of interest (NFI); centered on FI) was included. For models with instructional type, given that the category variable had three levels, the instructional type variable was also releveled to the segmental group as the baseline level in order to directly assess the difference between segmental and suprasegmental groups. This approach provides a direct test of group differences without impacting the goodness of fit of the model to the data (e.g. Linck & Cunnings, 2015).

All exploratory models were first run as forced entry models with maximal fixed effects and cross-classified subject and item intercepts (words nested within idiom for the logistic MEMs, and simply idioms for the ordinal MEMs). Random slopes were forward-tested one-by-one via likelihood ratio tests; only random slopes that significantly improved model fit and resulted in converging models were retained (Baayen, 2008; Baayen, Davidson & Bates, 2008). After the random effects structure was determined, backward testing of fixed effects was performed to arrive at the final model of best fit.

RESULTS

Intelligibility

Intelligibility was assessed with dictation accuracy. Learners' average accuracy, across all instructional groups, was just 34.6% (SD 17.4) at pretest and 42.9% (SD 18.2) at posttest. Such low accuracy indicates that listening to fluent speech was quite challenging for learners. Descriptive statistics for the dictation accuracy of each instructional group are reported in Table 1. In all the MEMs reported in the following sections, word category was significant ($p < .001$), and words that contained features of interest (FI) were transcribed with a much lower probability of an accurate response (45.6% – 52.0% lower, when averaging across test time and instructional type), indicating that the phonetic features taught did in fact make words less intelligible for learners.

<INSERT TALBE 1 ABOUT HERE>

TABLE 1

Descriptive Statistics for Intelligibility (SD)

Practice	Instruction	Pretest	Posttest
Production	Segmental	32.1% (19.0)	39.0% (16.3)
	Suprasegmental	31.5% (17.6)	38.0% (21.2)
Perception	Segmental	33.0% (14.4)	43.6% (16.5)
	Suprasegmental	30.4% (16.5)	43.2% (17.0)
None	None	37.4% (17.7)	45.0% (18.4)

PI (Both Types) with Production Practice. The model of best fit for PI followed by production-focused practice versus all controls, regardless of instructional type, is reported in

Appendix E. Participants given PI had significantly worse performance at pretest ($p = .021$). There was no time × instruction interaction, meaning that all participants improved at similar rates, with an improved probability of an accurate transcription by about 9.3% from pre to post ($p < .001$).

Segmental or Suprasegmental PI with Production Practice. The model of best fit for the production-focused practice groups versus all controls, with respect to instructional type, versus all controls was similar to the model comparing instruction generally to all controls, including the magnitude of the effect of test time and the lack of a time × instructional type interaction. The only new information revealed by the model is that the suprasegmental group ($b = -0.53$, $SE = 0.21$, $p = .010$) but not the segmental ($b = -0.34$, $SE = 0.13$, $p = .105$) was doing worse than the control group at pretest, and, upon releveling, the two instructional groups were not performing differently from each other ($b = -0.19$, $SE = 0.27$, $p = .480$).

PI (Both Types) with Perception Practice. The model of best fit for the PI followed by perception-focused practice versus all controls, regardless of instructional type, is reported in Appendix F. Participants given PI had significantly worse performance at pretest ($p = .036$). There was a time × instruction interaction such that those given PI performed better from pre to post than did the no instruction group ($p = .037$). The model estimates that, from pre to post, the group with no instruction improved by 10.1%, and those given PI improved by 13.7%.

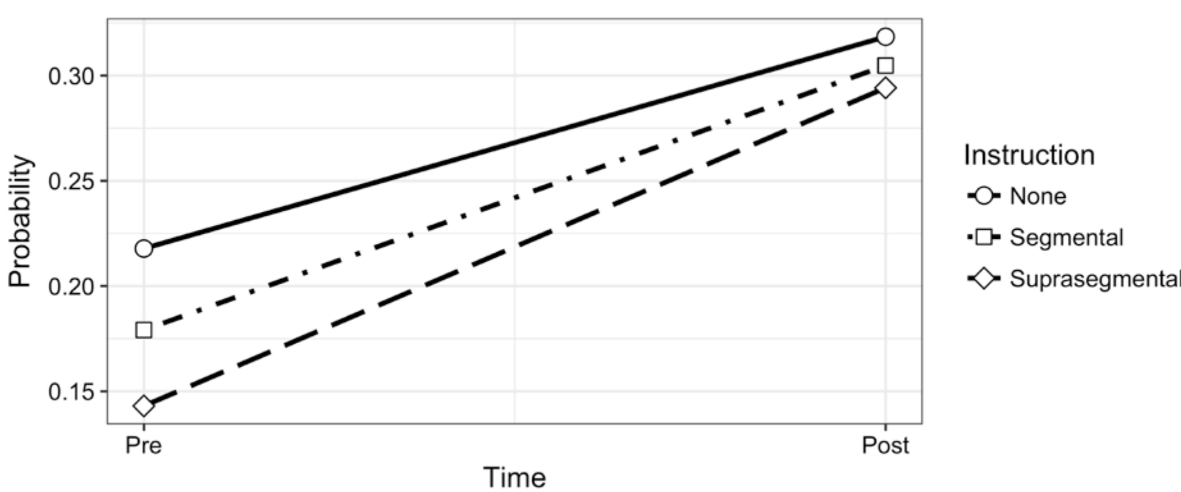
Segmental or Suprasegmental PI with Perception Practice. The model of best fit for the perception-focused practice groups versus all controls, with respect to instructional type, is reported in Appendix G and illustrated in Figure 1. The suprasegmental group ($p = .029$) but not the segmental group ($p = .249$) was significantly worse than the control group at pretest, and, upon releveling, the groups were not performing differently from each other ($b = -0.27$, SE =

0.28, *p* = .343). There was a time × instructional type interaction such that the suprasegmental group performed better from pre to post than did the no instruction group (*p* = .027). That is, even while controlling for the groups performing differently at pretest, the suprasegmental group improved to a significantly greater degree. Note the range of predicted performance for the model in Figure 1 showing that all participant groups were doing poorly overall and had room for improvement. Thus, it is not the case that one group performed worse at pretest and simply had more room for improvement. Finally, the interaction term for no instruction versus segmental instruction was not significant (*p* = .266) and, upon releveling, the term for segmental versus suprasegmental instruction was also nonsignificant (*b* = 0.22, *SE* = 0.22, *p* = .316). The model estimates that, from pre to post, the no instruction group improved by 10.1%, the segmental group by 12.6%, and the suprasegmental group by 15.1%. The effect for instruction versus no instruction observed above thus appears to be driven by the suprasegmental group.

<INSERT FIGURE 1 ABOUT HERE>

FIGURE 1

Probability of an Accurate Response in the Dictation Task (Intelligibility)



Comprehensibility

Comprehensibility was assessed with ratings of perceived difficulty per idiom, on a Likert scale of 1 – 7, with higher ratings indicating better comprehensibility. Descriptive statistics for the ratings are reported in Table 2. The low ratings indicate that learners found the fluent speech of the dictation task very difficult to process.

<INSERT TALBE 2 ABOUT HERE>

TABLE 2

Descriptive Statistics for Comprehensibility Ratings (SD)

Practice	Instruction	Pretest	Posttest
Production	Segmental	2.93 (1.49)	3.80 (1.50)
	Suprasegmental	3.01 (1.39)	3.27 (1.52)
Perception	Segmental	3.15 (1.14)	3.37 (1.54)
	Suprasegmental	2.97 (1.71)	2.99 (1.65)
None	None	2.77 (1.36)	2.98 (1.40)

PI (Both Types) with Production Practice. The model of best fit for PI followed by production-focused practice versus all controls, regardless of instructional type, is reported in Appendix H. Note that ordinal MEMs model thresholds for each Likert-scale level (1 – 7) separately. For ease of interpretation here, they were grouped based on the direction of effect, and those resultant groupings of the Likert scales are provided in the descriptions and graphs in the following sections.

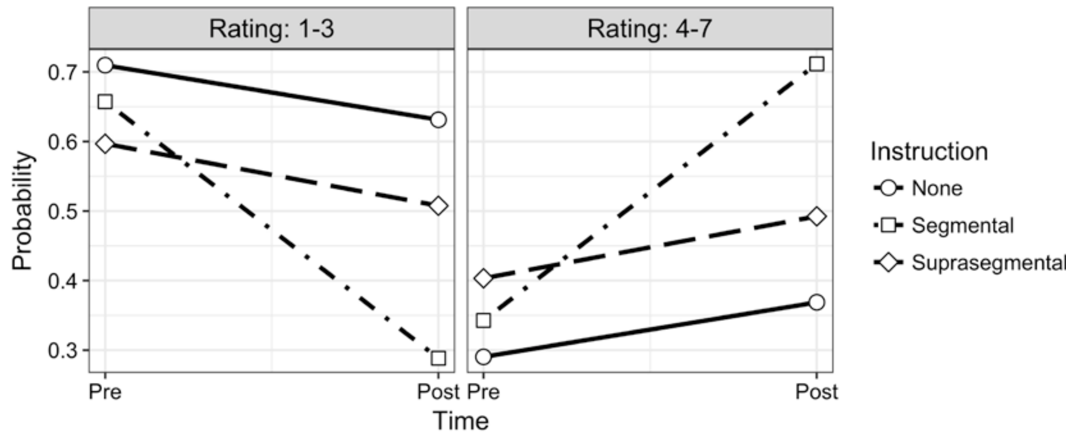
Participants given PI had significantly worse performance overall (p = .030), regardless of test time. There was no significant time × instruction interaction, but all participants reported better comprehensibility from pre to post (p < .001). Participants were 47.4% likely to rate comprehensibility at a 1 or 2 at pretest, but only 34.3% at posttest. Probability did not noticeably change for rating a 3 (23.1% – 23.8%). The probability of rating comprehensibility 4 – 7 increased from 29.5% to 42.0% at post. Overall, these results indicate that participants found the idioms more comprehensible over time.

Segmental or Suprasegmental PI with Production Practice. The model of best fit for PI followed by production-focused practice versus all controls, with respect to instructional type, is reported in Appendix I and illustrated in Figure 2. All changes in probability for ratings 1 – 3 from pre to post were negative and all those for ratings 4 – 7 were positive. There was a significant time × instruction interaction such that the segmental group found the idioms more comprehensible from pre to post than did the no instruction group ($p = .004$) and, upon releveling, also found the idioms more comprehensible pre to post than the suprasegmental group ($b = -1.19$, $SE = 0.52$, $p = .022$). The probably of rating an idiom 1 – 3 decreased by about from 65.7% to 28.8% for the segmental group pre to post whereas the no instruction group only decreased from 71.0% to 63.1% and the suprasegmental group from 59.7% to 50.8%. The probability of rating an idiom 4 – 7 increased pre to post from 34.2% to 71.2% for the segmental group and only from 29.0% to 36.7% and 40.3% to 49.2% for the no instruction and suprasegmental groups, respectively. It appears the effect of time observed was largely driven by the segmental group.

<INSERT FIGURE 2 ABOUT HERE>

FIGURE 2

Probability of Endorsing a Particular Comprehensibility Rating



PI (Both Types) with Perception Practice. After model testing, there were no significant effects of test time, instruction, or instructional type in any models with perception-focused practice. No model provided better model fit than the null models. Though their *p* values did not near significance, it is possible that the comprehensibility models lacked power compared to the intelligibility models due to having fewer observations; comprehensibility was assessed at the level of idiom (14 observations per participant), whereas intelligibility was assessed at the level of the word (96 observations).

The Effect of Comprehensibility on Intelligibility

Though intelligibility and comprehensibility are distinct constructs and so were measured and analyzed independently in this analysis, they are conceptually interrelated. In this study, a learner who believed an utterance to be incomprehensible might have chosen to transcribe few words from it and thus would receive a lower intelligibility score. A different speaker produced each idiom, and variation across speakers may have impacted comprehensibility. To account for this inter-speaker variability, idiom was included as a random effect in all the models reported thus far. However, to test directly the potential effect of comprehensibility on intelligibility, all models of intelligibility were run again, with comprehensibility ratings included as a fixed effect. The effect was significant ($p < .001$) in all models. The number of words students transcribed in idioms they rated as most comprehensible was 23.4% – 24.9% greater than in those they rated as the least comprehensible. This suggests that comprehensibility did have an impact on intelligibility for these listeners. However, the overall pattern of results—i.e. the instructional group and practice effects reported—did not change when comprehensibility was included in the models.

DISCUSSION

Prior studies have found that pronunciation instruction (PI) can improve L2 learners' global pronunciation in terms of intelligibility, comprehensibility, and/or accent (Lee, Jang & Plonsky, 2015) as long as enough time is allotted for PI (Saito, 2012). Prior studies have also found that intensive exposure and practice to difficult sounds can improve L2 learners' perception (e.g. Thomson, 2012) and that such training can also impact production (Sakai & Moorman, 2018). Theories of L2 phonological acquisition posit a close connection between perception and production (Flege, 1995). This study sought to determine whether the purported benefit of classroom PI extends to perception under more real-world conditions, listening to fluid target language speech in sentence-level utterances, for relatively novice learners. The study sought to find out if PI makes target language speech more intelligible or comprehensible for beginning L2 learners (RQ1) and if so, under what types of practice conditions—perception-focused or production-focused practice—(RQ2), and with what type of feature target—segmental or suprasegmental (RQ3). At first glance a *pronunciation* intervention to improve listening might seem to turn the predictions of the Speech Learning Model (Flege, 1995) on their head, since the model proposes that accurate perception is a necessary precursor to accurate production. However, it was hypothesized that since PI seems to attune learners to sounds in their L2, what they acquire in PI could be recruited to support bottom-up processing during listening and thus improve the intelligibility and comprehensibility of target language speech for L2 listeners. Beginning SFL learners were exposed to 1.3 hours of PI. They demonstrated a small but significant improvement in listening skills as assessed by a dictation task. PI improved the intelligibility and comprehensibility of target language speech for these learners, but the effect varied across outcome measure, focus of practice, and features targeted. PI improved intelligibility (measured as number of words correctly transcribed) when it was accompanied by perception-focused practice and when it targeted suprasegmental features. The pattern of results was quite different for comprehensibility.

Intelligibility

L2 learners have difficulty locating word boundaries and identifying words embedded in the speech stream because in the early stages of learning, they rely on heuristic procedures that are initially dictated by the structure of their L1 (Cutler, 2001). Learners tend to activate 'phantom' words that are phonologically plausible but not existent in the L2 (Broersma & Cutler, 2008), allow top-down expectations to overrule bottom-up information detected (a 'lexical effect'), and be reluctant to revise their predictions when confronted with conflicting evidence in the signal (Field, 2004). The learners in this study demonstrated all these inefficient processing routines. For instance, many substituted the familiar words *buenos hombres* 'good men' for the less familiar sequence *buena sombra* 'good shade.' Similarly, they substituted *vistos* 'seen' for *vicios* 'vices.' Learners extracted pseudo words like *cabar* from the sequence *va a acabar* 'is going to end.' Instead of *que el* 'that the' they heard the pseudo word *quel.* As a result, they found the fluent speech of the dictation task largely unintelligible, correctly transcribing only a third of it in the pretest.

PI targeting suprasegmental features, with perception-focused practice, helped learners segment the speech stream and identify known and unknown words in the dictation task. For instance, in the posttest, after learning about synalepha (Appendix A, Module 4), learners could identify three words in the phrase *limpia el trigo* ‘clean the wheat,’ which features synalepha, whereas in the pretest they more often heard *limpia trigo* ‘clean wheat’ or *limpiar trigo* ‘to clean wheat.’ After learning about identical vowels in contact, learners could identify word boundaries in the phrases *que el* and *va a acabar*, which are monosyllabic and trisyllabic phrases, respectively, in fluent speech. Thus, suprasegmental PI appears to have helped learners reduce phantom word activation, lexical effects, and use of top-down expectations generally, facilitating more accurate bottom-up processing largely in terms of speech segmentation.

It was hypothesized that segmental PI would be facilitative of another aspect of bottom-up processing, which is word identification. For instance, in the pretest many learners heard the cognate *castillo* ‘castle’ rather than the less familiar word *castigo* ‘punishment.’ It was expected that learning about the velar approximant allophone of /g/ in *castigo* would help them not mistake it for the palatal fricative of *castillo*. Similarly, it was expected that learning about the alveolar tap and the approximant allophone of /d/ would help them correctly identify *cordero* ‘lamb’ rather than mishear it as *correo* ‘mail’ as they did in the pretest. Indeed, learners who received segmental PI did correct many of these errors, but not at rates significantly higher than the learners who did not receive PI. All groups improved over time at roughly equal rates, suggesting that the input typical of communicative classrooms was sufficient to improve segmental perception, at least with the items measured in this dictation task.

The dictation task featured equal numbers of segmental and suprasegmental target features, but the suprasegmental PI group improved more than others on this task over time. The results suggest that PI targeting suprasegmental features is beneficial, and segmental PI seems relatively more dispensable. However, practice is key; the PI groups that only practiced pronunciation did not find the dictation task speech significantly more intelligible than the control group, whereas learners who got perception practice with feedback did. Of course, pronunciation practice is an essential component of PI, and it was included here. The added advantage gleaned by the learners who also received perception-focused practice suggests that learners can be taught how to transfer their burgeoning knowledge of the L2 sound system to the perceptual domain. In many other areas learners seem to require practice to apply knowledge to a new context or domain, and it is logical that this be true for phonology as well.

Importantly, though, the significant group differences reported here seem attributable to instruction followed by practice and feedback, not merely practice alone. Several features of the study design make it implausible that the results could be explained simply as practice with the task. First, the total practice time allotted was brief (10 minutes per target). Second, the time elapsed between pre and posttest (12 weeks) was sufficient to preclude a task-specific practice effect, because learners could not retain the pretest in their memory for three months. Third, the task included many unfamiliar words that were not included in any instructional materials or practice (e.g. *trigo*, *acabar*, *vicios*). Fourth, the task presented learners with speech from a variety of speakers and dialects, none of which learners heard during instruction or practice.

Comprehensibility

The benefit of suprasegmental PI for intelligibility was not, however, evidenced for comprehensibility. On the contrary, the group that received *segmental* PI and *pronunciation* practice found the idioms in the dictation more comprehensible over time. Comprehensibility is an impressionistic measure of how difficult a rater/listener believes it is to understand an utterance. That judgment may or may not impact intelligibility, which is how much a rater/listener actually understands. Previous studies have found that PI can affect the comprehensibility (as well as accentedness) of L2 learners' speech, as rated by native speakers of the target language (e.g. Derwing, Munro & Wiebe, 1998). The current study employed comprehensibility judgments in the opposite direction, asking L2 learners to rate how comprehensible they believed target speech to be. Given the relative paucity of research on the subject, no strong prediction was made, but the perhaps surprising results bear further discussion.

There are several possible explanations for why instructional effects patterned differently across outcome measures, focus of practice, and instructed features. First, the results could be an artifact of the particular methods used here (e.g. proficiency level, target features, practice procedures, dictation task) or the fact that the comprehensibility models lacked power compared to the intelligibility models due to having fewer observations and so were unable to reveal patterns that existed. Replication studies with larger sample sizes could attempt to tease out these possibilities. Another interesting possibility is that learners' beliefs about their learning impacted their comprehensibility judgments in unanticipated ways. Though all learners evaluated the PI favorably on a post-instructional questionnaire, a simple (non-statistical) comparison of raw scores suggested that those who received segmental PI believed it was slightly more helpful (average of 4.08 out of 5), interesting (4.08), and facilitative of learning generally (4.23) and listening specifically (3.62) compared to the suprasegmental PI group. That group gave slightly lower evaluations in helpfulness (3.83), interest (3.92), learning (4.00), and listening improvement (3.46). Most student comments were neutral to favorable, but a few comments from the segmental PI group stood out as indicating how relevant they found PI, such as one student's report that "It's probably the best thing we did to improve listening comprehension." On the other hand, the suprasegmental PI group's comments indicated more uncertainty about their ability to apply what they learned to the dictation task, for instance, “They [the lessons] were helpful, but it’s still hard to comprehend other dialects that we’re not used to hearing.” These beliefs could have led learners in the segmental PI group to be more aware of their progress over time or to feel more confident as they took the posttest, increasing comprehensibility ratings. Another possible explanation is that word identification and word segmentation are processes that learners experience as qualitatively different. They might, for instance, be aware of when they struggle to identify particular phonetic segments in a word and thus have to guess about which word they heard (e.g. *castigo* vs. *castillo*), whereas they might be relatively less cognizant of when they cannot segment a phrase and thus fail to hear some words entirely (e.g. *va a acabar*). If this is true, then learners who received segmental PI might notice their improved ability to correctly identify unknown words based on specific target segments in a way that the suprasegmental PI group would not. Likewise, learners’ beliefs about what constitutes relevant practice (perception or production) could have impacted their comprehensibility judgments, perhaps leading the production-practice group to become more confident in listening to the target language. Though plausible, all these hypotheses require direct testing in future studies.

The results here suggested some relationship between comprehensibility and intelligibility of target language speech for L2 listeners, since they transcribed fewer words on idioms that they judged to be more difficult to comprehend than on idioms they judged to be less difficult. However, the group that judged target language speech to be more comprehensible over time (segmental instruction + production practice group) did not actually find the speech more intelligible, as measured by number of words transcribed on the dictation task. The group that found improved intelligibility (suprasegmental instruction + perception practice) did not judge the language speech to be more comprehensible over time. Thus, given the overall pattern of results, this study contributes to the growing body of research pointing to the separability of these two constructs—intelligibility and comprehensibility—in L2 phonological acquisition.

Intelligibility is arguably more important because it objectively measures learners’ ability to perform an important real-world task, which is to process target language speech, whereas comprehensibility is an impressionistic judgment that may have little bearing on intelligibility. One might wonder if comprehensibility judgments are relevant at all. In the opinion of the author, comprehensibility matters because it matters to learners themselves. They want to feel more accomplished as listeners. In turn, their feelings about both their developing abilities in the L2 and about the relevance of instruction for their learning could determine the degree to which they engage with the target language and the instruction. Thus, comprehensibility has the potential to impact development of listening skill over the longer term. Therefore, even though segmental PI did not have a clear impact on intelligibility in this study, a holistic and cautious interpretation of the results suggests that there is still reason to include segmental features and pronunciation practice in a PI intervention aimed at improving L2 learners’ bottom-up processing of the target language speech stream.

CONCLUSION

This study sought to improve L2 learners’ bottom-up processing during listening by exposing them to PI that would make them more aware of several difficult-to-perceive features of the target language sound system. The study found that PI on suprasegmental features, along with perception-focused practice, prompted learners to more accurately segment target language speech and find it more intelligible. PI on segmental features did not have the same effect, though learners who received segmental PI with production-focused practice judged the target language speech to be more comprehensible.

This study makes several contributions to the growing body of literature on PI. First, the participants in the study were relative novices, whereas prior studies of PI in the SFL context have more often recruited learners enrolled in advanced phonetics courses. It was thought that the current intervention was best suited for the novice level because novices rely more on bottom-up processing while listening. Second, though perception and production are thought to be intimately related, they are usually taught and tested separately in the research. Many classroom studies of PI have measured changes in learners' pronunciation, and laboratory studies of perceptual training have measured changes in learners' perception, but this study investigated the crossover effect of PI on perception. Finally, both laboratory training studies and classroom studies often employ highly controlled tasks with very short segments of speech in isolation. In this study learners listened to fluent sentences from a variety of speakers, testing the potential effect of PI in a task with relatively more face and ecological validity.

The study has clear pedagogical applications. Namely, PI can benefit leaners in more ways than pronunciation. Teachers who want to provide bottom-up processing instruction for listening should consider incorporating PI, even if their instructional time for PI is limited. The PI should target a range of features and focus on suprasegmental features that differ between the L1 and L2. The PI should include production practice with feedback and also perception practice with feedback to help learners transfer their bottom-up processing skills to the perceptual domain. The perception practice might take the form of dictation tasks.

Dictation tasks do, however, present some methodological challenges, as discussed at length by Yeldham (2017). The task used here included unknown words embedded in sentence-level discourse and required full transcription. A different task or different conditions may have produced different results. Prior studies have reported that suprasegmental awareness is more important for less controlled tasks (e.g., Gordon & Darcy, 2016), so the task employed here for sake of face and ecological validity may have advantaged the suprasegmental group. Future studies should investigate the effect of the assessment task. Future studies should also investigate a greater variety of segmental and suprasegmental features in PI, as the ones selected here might not be the most important features for bottom-up processing, and should examine the effect of instruction over a longer period of time.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to all the students who participated in the study, the language program directors and instructors who supported the project, and the speakers who provided the stimuli. Special thanks go to Michele Mauney for assistance with transcription, Nick Pandža for assistance with statistical analyses, and to my fellow linguistics colleagues at the University of Richmond and anonymous reviewers for providing helpful suggestions. Any remaining errors are my own. The University of Richmond's School of Arts and Sciences generously provided summer research fellowships to fund this project

REFERENCES

Ausín, A., & Sutton, M. (2010). An L2 pronunciation judgment task. In C. Borgonovo et al. (Eds.), *Selected Proceedings of the 12th Hispanic Linguistics Symposium* (pp. 234–245). Somerville, MA: Cascadilla Proceedings Project.

Akerberg, M. (2005). La percepción auditiva como factor en la adquisición de sonidos de L2 [Auditory perception as a factor in the acquisition of L2 sounds]. In M. Akerberg (Ed.), *Adquisición de segundas lenguas: Estudios y perspectivas* [Second language acquisition:

Studies and perspectives] (pp. 45–72). Mexico City, Mexico: Universidad Nacional Autónoma de México.

Baayen, H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, UK: Cambridge University Press.

Baayen, H., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory & Language, 59*(4), 390–412.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: Linear mixed-effects models using eigen and S4. R package version 1.1-9. Accessed 6 March 2018 at: http://CRAN.Rproject.org/package=lme4

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

Broersma, M., & Cutler, A. (2008). Phantom word activation in L2. *System, 36*(1), 22–34.

Brown, S. (2013). *Listening myths: Applying second language research to classroom teaching*. Ann Arbor, MI: The University of Michigan Press.

Broselow, E., & Park, H. B. (1995). Mora conservation in second language prosody. In J. Archibald (Ed.), *Phonological acquisition and phonological theory* (pp. 151–168). Hillsdale, NJ: Lawrence Erlbaum.

Christensen, R. H. B. (2015). Ordinal regression models for ordinal data. R package version 2015.6–28. Accessed 6 March 2018 at: http://www.cran.r-project.org/package=ordinal/

Colantoni, L., & Steele, J. (2008). Integrating articulatory constraints into models of second language phonological acquisition. *Applied Psycholinguistics, 29*(3), 489–534.

Cutler, A. (2001). Listening to a second language through the ears of a first. *Interpreting, 5*(1), 1–23.

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39*(3), 379–397.

Field, J. (2003). Promoting perception: Lexical segmentation in L2 listening. *ELT Journal, 57*(4), 325–334.

Field, J. (2004). An insight into listeners' problems: Too much bottom-up or two much top-down? *System, 32*(3), 363–377.

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

Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software, 8*(15), 1–27.

Goh, C. C. M. (2000). A cognitive perspective on language learners' listening comprehension problems. *System, 28*(1), 55–75.

Gordon, J., & Darcy, I. (2016). The development of comprehensible speech in L2 learners. *Journal of Second Language Pronunciation, 2*(1), 56–92.

Graham, S. (2006). Listening comprehension: The learners' perspective. *System, 34*(2), 165–182.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, *38*(2), 201–223.

Huensch, A. (2016). Perceptual phonetic training improves production in larger discourse contexts. *Journal of Second Language Pronunciation*, *2*(2), 183–207.

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

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

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

Kewley-Port, D., Nishi, K. & Park, H., Miller, J., & Watson, C. (2009). Learn to Listen (L2L): Perception training system for learners of English as a second language. *The Journal of the Acoustical Society of America, 125*(4), 2773.

Kilpatrick, C. D., & McLain Pierce, L. (2014). Focused instruction in Spanish syllabification. *Foreign Language Annals, 47*(2), 286–299.

Kissling, E.M. (2015). Phonetics instruction improves learners' perception of L2 sounds. *Language Teaching Research, 19*(3), 254–275.

Kissling, E. M. (2013). Teaching pronunciation: Is explicit phonetics instruction beneficial for FL learners? *The Modern Language Journal, 97*(3), 720–744.

Leather, J. H., & James, A. R. (1996). Second language speech. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 269–316). New York: Academic Press.

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

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

Levis, J. M. (2016). The interaction of research and pedagogy. *Journal of Second Language Pronunciation, 2*(1), 1–7.

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

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning, 65*(S2), 185–207.

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

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and / l /: A first report. *Journal of the Acoustical Society of America, 89*, 874–866.

Logan, J. S., & Pruit, J. S. (1995). Methodological issues in training listeners to perceive non-native phonemes. In W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research* (pp. 351–377). Timonium, MD: York Press.

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

Nishi, K., & Kewley-Port, D. (2008). Nonnative speech perception training using vowel subsets: Effects of vowels in sets and order of training. *Journal of Speech, Language, and Hearing Research, 51*, 1480–1493.

R Core Team (2015). R: A language and environment for statistical computing. Vienna: R

Foundation for Statistical Computing. Accessed 6 March 2018 at: http://www.R-project.org/

Rasmussen, J., & Zampini, M. (2010). The effects of phonetics training on the intelligibility and comprehensibility of native Spanish speech by second language learners. In J. Levis & K. LeVelle (Eds.), *Proceedings of the 1st Pronunciation in Second Language Learning and Teaching Conference*, Iowa State University, September 2009 (pp. 38–52). Ames, IA: Iowa State University.

Saito, K. (2012). Effects of instruction on L2 pronunciation development: A synthesis of 15 quasi-experimental intervention studies. *TESOL Quarterly, 46*(4), 842–854.

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

Shin, D. J., & Iverson, P. (2013). Training Korean second language speakers on English vowels and prosody. *Proceedings of Meetings on Acoustics, 19*, 1–4.

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

Siegel, J., & Siegel, A. (2015). Getting to the bottom of L2 listening instruction: Making a case for bottom-up activities. *Studies in Second Language Learning and Teaching, 5*(4), 643–662.

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

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

Vandergrift, L., & Baker, S. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning*, *65*(2), 390–416.

Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and learning second language listening: Metacognition in action.* New York: Routledge.

Wilson, I., Kaneko, E., Lyddon, P., Okamoto, K., & Ginsburg, J. (2011). Nonsense-syllable sound discrimination ability correlates with second language (L2) proficiency. In W. S. Lee & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 2133–2136). Hong Kong: City University of Hong Kong.

Yeldham, M. (2016). Second language listening instruction: Comparing a strategies-based approach with an interactive, strategies/bottom-up skills approach. *TESOL Quarterly, 50*(2), 394–420.

Yeldham, M. (2017). Techniques for researching L2 listeners. *System*, *66*(1), 13–26.