To be published in *Language Learning* (Wiley Blackwell)

LANGUAGE LEARNING

A Journal of Research in Language Studies

Effects of Second Language Pronunciation Teaching Revisited: A Proposed Measurement Framework and Meta-Analysis

Kazuya Saito & Luke Plonsky

Abstract

We propose a new framework for conceptualizing measures of instructed L2 pronunciation proficiency according to three sets of parameters: (a) the constructs being focused on (global vs. specific), (b) the scoring method (human raters vs. acoustic analyses), and (c) the type of knowledge being elicited (controlled vs. spontaneous). Adopting the model (i.e., Framework for L2 Pronunciation Measurement) as a synthetic tool, we code the instruments found in 77 studies of L2 pronunciation teaching published between 1982 and 2017. We calculate the frequency of each measurement type and re-examine the interaction of instructional effectiveness and measurement within the sample. According to the results, instruction is most effective when it targets learners' monitored production of specific segmental/suprasegmental features. The efficacy of instruction remains relatively unclear when gains are measured globally via subjective/human judgements especially at a spontaneous level. The findings are discussed to improve the designs in L2 pronunciation research and, more generally, strengthen the interface between pronunciation teaching, measurement and SLA.

Key words: Second language pronunciation; pronunciation teaching; instructed SLA; meta-analysis; research synthesis

Acknowledgement

We are grateful to five *Language Learning* reviewers for their insightful comments on earlier versions of our manuscript and to Associate Editor, ***Emma Marsden*** for her advice, guidance and dedication at every stage of the review process.

Introduction

Since publication of the special issue “Changing Contexts and Shifting Paradigms in Pronunciation Teaching” in *TESOL Quarterly* (Levis, 2005), which featured a number of position and empirical papers by key scholars, a recognized paradigm shift has been stimulated in the field of second language (L2) pronunciation research. Many of these scholars have stressed the importance of explicit pronunciation instruction as a way to help L2 learners attain “intelligible” pronunciation, but not as a de-contextualized, nativist form of language development (reminiscent of audio-lingualism). Over the past 10 years, the number of pronunciation teaching studies has dramatically increased with numerous promising findings, indicating that explicit instruction—operationalized as provision of articulatory (how to produce) or/and auditory (how to hear) information about L2 segmental and suprasegmental features—can impact adult L2 learners’ pronunciation proficiency to a great degree (e.g., Derwing, Munro, Foote, Waugh, & Fleming, 2014; Kissling, 2013; Trofimovich, Kennedy, & Blanchet, 2017). Lee, Jang and Plonsky’s (2015) meta-analysis of 86 studies revealed a medium effect of L2 pronunciation instruction relative to comparison groups which did not receive pronunciation-focused instruction ($d = 0.80$; 95% CIs [.77, .81]). Focusing on 18 perception training studies, Sakai and Moorman’s (2018) meta-analysis found that those who received explicit instruction demonstrated medium-sized improvement in not only their perception ($d = 0.93$; $SD = .72$), but also production abilities ($d = 0.89$; $SD = .61$) compared to those who did not.

In their narrative review, however, Thomson and Derwing (2015) pointed out that most of the existing literature appeared to have an implicit conceptual focus on the mastery of nativelike pronunciation, with monolingual native speakers being the ideal model and goal. The dominance of the nativism paradigm is surprising, given that there is ample research evidence that few adult L2 learners can attain such accent-free speech (e.g., Flege, Munro, & MacKay, 1995), and that many L2 speech researchers have stressed time and time again that L2 students should be encouraged to pursue more realistic and achievable goals, such as the acquisition of intelligible and comprehensible pronunciation despite detectable accent (Isaacs, Trofimovich, & Foote, 2017; Levis, 2005; Munro & Derwing, 1995). Accordingly, pronunciation teaching should rather be designed to help L2 students acquire what matters for their future real-life L2 usage in the most efficient and effective way.

In the current study, apart from addressing the conceptual underpinnings of pronunciation teaching (i.e., the intelligibility vs. nativeness principles: Levis, 2005), we would like to elucidate another set of crucial issues in the pronunciation teaching literature strongly related to theory and practice: the match between what pronunciation instruction intends to teach, how its effectiveness is assessed, and how the resulting outcomes are interpreted in light of current theories and relevant discussion. As both meta-analytic (Lee et al., 2015) and narrative (Thomson & Derwing, 2015) reviews have noted, the pedagogical efficacy of pronunciation teaching in primary studies has been scrutinized from multiple angles using a wide variety of tasks and scoring methods. These outcome measures greatly differ according to the types of features being targeted (global comprehensibility and intelligibility vs. specific segmental and suprasegmental measures) and the scoring methods that are employed to measure them (e.g., impressionistic judgements vs. acoustic analyses). In addition, whereas some tasks elicit controlled production (e.g., word and sentence reading), others involve more spontaneous speech (e.g., picture description). There is no single best assessment method; rather, each instrument carries with it a set of trade-offs (e.g., control over phenomena under investigation vs. ecological validity). Unfortunately, researchers who study pronunciation teaching often apply different measures idiosyncratically, and without apparent consideration of the impact that different choices in measurement may have on performance or study outcomes (e.g., Norris & Ortega, 2012).

The lack of an overarching model for measuring the effects of instruction or for L2 pronunciation more generally constrains the development of constructive discussion on the interaction of type of instructional treatment, pronunciation phenomena, and resulting knowledge. Whereas it is well-accepted that pronunciation teaching can make a difference in some forms of L2 pronunciation learning, it has remained surprisingly unclear the extent to which different aspects of L2 pronunciation proficiency and knowledge are actually improved by such interventions. It is also unclear how the choice of pronunciation measures may influence our ability to detect changes resulting from instruction.

As a remedy, the current study aims to achieve two objectives. By incorporating well-established insights in instructed second language acquisition (SLA) (e.g., DeKeyser, 2017 for Skill Acquisition Theory) and L2 pronunciation (e.g., Flege, 2016 for Speech Learning Model; Major, 2008 for Ontogeny Phylogeny Model) into the context of pronunciation teaching research, we will first propose a unified model that future researchers can use to measure the effect of instruction on L2 pronunciation development in a more systematic and principled manner—i.e., the Framework for L2 Pronunciation Measurement. To this end, we will carefully address the crucial question of how to define the various constructs of L2 pronunciation *proficiency* according to global (comprehensibility, intelligibility, perceived fluency, accentedness) vs. specific (segmentals, suprasegmentals) distinctions, and different types of subjective (impressionistic judgements) and objective (acoustic analyses) scoring methods. Subsequently, we will elaborate a task taxonomy by which to measure the impact of instruction on the development of L2 learners' pronunciation *knowledge* under different conditions (such as controlled or more spontaneous tasks). The hierarchy of the framework we propose here is visually summarized in Figure 1.

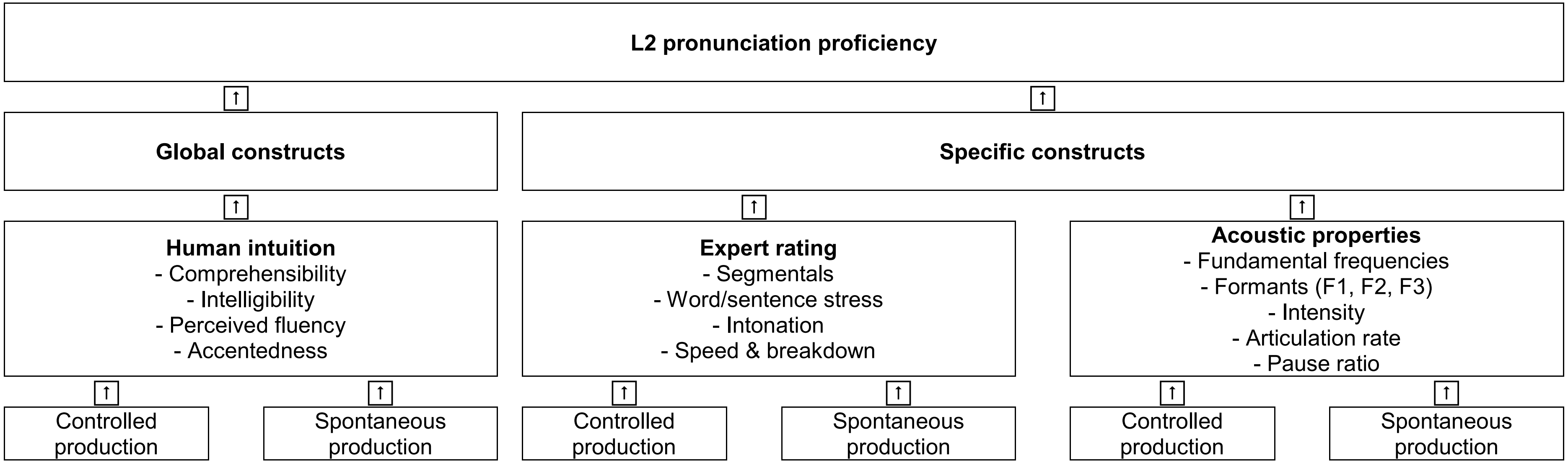

Figure 1. Framework for L2 Pronunciation Measurement: A Proposed Model for Measuring the Effectiveness of L2 Pronunciation Teaching

Here, we acknowledge that our model of proficiency and knowledge is exploratory; and our assumption that different types of analyses and tasks reflect different types of L2 pronunciation proficiency and knowledge is suggestive at best. To demonstrate the extent to which the Framework for L2 Pronunciation Measurement can actually serve as a useful ontology, the second objective of the study will conduct a research synthesis and meta-analysis of *N* = 77 quasi-experimental studies published in international journals between 1982 and 2017. First, the proposed model will be used to re-examine the research practices and findings in pronunciation teaching literature. We will survey how primary studies have measured instructed L2 pronunciation learning as per different constructs of measurement focus (global vs. specific), scoring method (subjective vs. objective), and task type (controlled vs. spontaneous).

Finally, the finding of the primary studies will be meta-analyzed by incorporating measurement focus (global vs. specific), scoring method (subjective vs. objective) and task type (controlled vs. spontaneous) as moderator variables. Here, we aim to test our prediction that instruction is facilitative of L2 pronunciation development with the amount of effectiveness being tied to the type of outcome measures specified in our proposed model. Through achieving these objectives (proposing, elaborating and applying the measurement framework of instructed L2 pronunciation), it is believed that the model will help future pronunciation teaching researchers bring to light the pedagogical and theoretical relevance of their primary studies in a more empirically-grounded, straightforward and accessible fashion. What we ultimately aim to provide with this model and these results is a framework on which pronunciation teaching researchers can base their choices about how to measure and interpret the effects of instruction.

Different Constructs of L2 Pronunciation Proficiency

Few pronunciation teaching researchers would disagree with the fundamental idea that the ultimate goal of pronunciation teaching is to improve L2 pronunciation proficiency in terms of enhanced intelligibility and comprehensibility. As we will show in the research synthesis presented later, however, many primary studies of pronunciation teaching have failed to provide detailed explanation on the following logical connection: Whether, to what degree, how and why the focus of instruction—i.e., the improvement of individual segmental and suprasegmental features—can lead to the development of overall L2 pronunciation proficiency (e.g., comprehensibility). To this end, we need clear definitions for L2 pronunciation proficiency according to different types of analysis and scoring methods adopted in primary studies.

In our proposed framework, L2 pronunciation proficiency is considered as a multilayered phenomenon composed of three different levels—one global construct (human intuition of overall pronunciation proficiency) and two specific constructs (expert rating of segmental and suprasegmental pronunciation proficiency and computerized measurement of acoustic properties) (see the second and third rows of Figure 1). We argue that the three different constructs of L2 pronunciation proficiency need to be carefully distinguished so that future pronunciation teaching researchers can clarify the relationship between the specific target of instruction (segmental and suprasegmental features), outcome measures (pre/posttests) and the subsequent impact on overall L2 pronunciation proficiency (e.g., comprehensibility).

Human Intuitions of Global Proficiency

L2 pronunciation proficiency is first captured as a part of global L2 oral proficiency, and is typically measured through native (and non-native) listeners' holistic judgements of L2 talkers' speech samples, which are generally played only once without any detailed descriptors or training. A crucial characteristic of measuring this global proficiency (whether it involves rating or transcription) is that researchers provide listeners with very brief definitions of each construct that listeners must rate without any explicit mention of certain pronunciation features. This is because this construct is believed to provide insights into the quick, intuitive assessments of L2 use that are typical of real-life contexts. To date, scholars have used several listener-based global measures to assess L2 speech, such as comprehensibility (i.e., ease of understanding), perceived fluency (i.e., the flow and smoothness of speech), and accentedness (i.e., linguistic nativelikeness). Another listener-based intuitive approach is intelligibility (i.e., actual product of understanding), wherein listeners are typically asked to transcribe what they have heard from each speech sample (for a list of other types of intelligibility measures [e.g., responses to true/false statements, comprehension questions, scalar ratings], see Munro & Derwing, 2011).

Our definition of global L2 pronunciation proficiency—i.e., listeners' holistic judgements of L2 speech—is specific to the context of pronunciation teaching research. The definition is developed to capture a range of measures that primary studies have adopted for assessing the effectiveness of instruction (e.g., comprehensibility, intelligibility, perceived fluency, accentedness). However, we do acknowledge that what characterizes global aspects of L2 pronunciation proficiency has remained open to debate in the field of L2 assessment (for a review, see Harding, 2017). To further examine this topic, some scholars have attempted to align what pronunciation teaching researchers are primarily concerned with (e.g., L2 comprehensibility) with major pronunciation proficiency benchmarks (e.g., IELTS Pronunciation Scale) (Issacs, Trofimovich, Yu, & Muñoz Chereau, 2015).

In our model, listeners' ratings (comprehensibility, accentedness, perceived fluency) and transcriptions (intelligibility) are clustered as a single construct (i.e., global L2 pronunciation proficiency), but distinguishable from other types of specific L2 pronunciation proficiency (i.e., expert raters' scalar judgements and acoustic analyses of segmentals and suprasegmentals). This methodological distinction is in line with empirical evidence in existing L2 pronunciation research. For example, significantly strong correlations ($r > .80$) have been observed between comprehensibility, accentedness and intelligibility (e.g., Munro & Derwing, 1995); and comprehensibility and perceived fluency (e.g., Derwing, Rossiter, Thomson, & Munro, 2004). The findings here suggest that these global measures could be considered as a *statistically* similar phenomenon.

Comparatively, the strength of correlation coefficients between global L2 pronunciation proficiency (overall comprehensibility, accentedness, perceived fluency of speech) and specific L2 pronunciation proficiency (segmental, prosodic and temporal features of speech) widely varies in accordance with types of constructs of interest (*r* = .1-.8). For example, listeners' global accentedness judgements were strongly tied to segmental accuracy (*r* = .80 in Riney, Takada, & Ota, 2000), and they appeared to have only weak associations with, in particular, melody-based characteristics of L2 speech (e.g., *r* = .15 for peak alignment in Trofimovich & Baker, 2006). The results of these studies here suggest that global and specific constructs of L2 pronunciation may be statistically independent constructs to some extent, because the relationship between global and specific L2 pronunciation proficiency appears to be non-linear. It could be considerably strong for some features (e.g., accentedness vs. segmental accuracy), but may become relatively weak for other features (e.g., accentedness vs. melody-based suprasegmentals).

As evidenced in the previous literature (e.g., Derwing, Munro, & Wiebe, 1998), teaching individual segmental and suprasegmental features can positively influence the global construct of L2 pronunciation proficiency. This is because much of the variance in human intuitions of L2 comprehensibility, fluency and accentedness ratings (e.g., 60-70%) is primarily accounted for by the phonological qualities of speech (Crowther, Trofimovich, Saito, & Isaacs, 2015; Isaacs & Trofimovich, 2012; Kang et al., 2010; Saito, Trofimovich, & Isaacs, 2017). Notably, scholars have also identified other factors that may explain a substantial portion of remaining construct-relevant (i.e., non-error) variance, such as other linguistic elements (e.g., Saito, Webb, Trofimovich, & Isaacs, 2016 for lexicogrammar accuracy, fluency and complexity) and listener backgrounds (e.g., Kennedy & Trofimovich, 2008 for familiarity with particular accents; Saito et al., in press for L1 vs. L2 listeners). In this regard, the impact of instruction on global L2 pronunciation proficiency is likely to be partial at best; even if the pronunciation teaching focuses on specific phonological features, L2 learners' improved global pronunciation proficiency will also depend on other non-phonological factors as well.

This conceptual discussion about the relationship between global and specific L2 pronunciation proficiency is summarized in Figure 2.

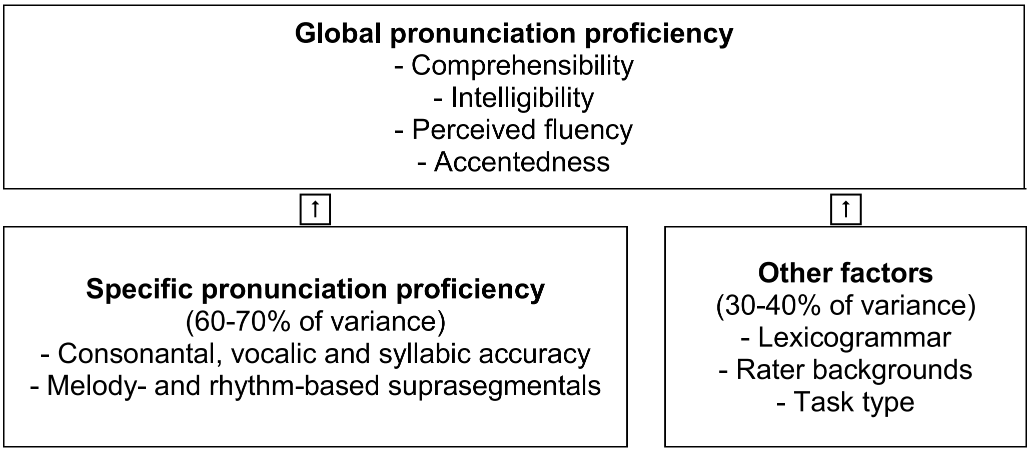

Figure 2. A Visual Summary of Global and Specific L2 Pronunciation Proficiency

Expert rating of Specific Proficiency

When it comes to human judgments that involve global pronunciation proficiency ratings, non-expert "listeners" with varying degrees of familiarity with foreign accented speech are typically recruited (Kennedy & Trofimovich, 2008). In the pronunciation teaching literature, a number of specific features or constructs of L2 pronunciation proficiency have also been defined and judged by expert raters. These include learners' capacities to (a) pronounce new consonantal and vocalic sounds in an L2 without deleting/substituting them for first language counterparts in both simple (e.g., Consonant-Vowel [CV]) and complex (e.g., CVC, CCVCC) syllable structures (known as segmental and syllabic accuracy, respectively); (b) use adequate prosody at the word (correct assignment of word stress) and sentence (e.g., appropriate use of intonation for declarative and interrogative intensions) levels; and (c) deliver speech at an optimal tempo (speed fluency) without taking too many pauses (breakdown fluency) or making too many repetitions/self-corrections (repair fluency). According to Trofimovich and Baker's (2006) framework, acoustic phenomena signaled via fundamental frequency and intensity (word stress, intonation) are categorized as "melody-based" suprasegmentals; and all the temporal characteristics (speed, breakdown, repair fluency) are referred to as "rhythm-based" suprasegmentals.

To evaluate such specific aspects of L2 pronunciation proficiency, researchers have recruited and trained expert (rather than naïve) raters with a great amount of linguistic and pedagogical experience (Isaacs & Thomson, 2013). The assessment of specific proficiency features requires experienced "raters" with teaching and/or linguistics backgrounds due to the demanding nature of the tasks (i.e., evaluating only phonological qualities of speech) (for further discussion and justification on the distinction between listeners vs. raters, see Yan & Ginther, 2017).

In this approach, expert raters carefully listen to speech samples (multiple times) and then analyze *only* the phonological qualities of the samples, such as segmentals, word stress, intonation and speech rate (Saito et al., 2017). In other studies, expert raters make accuracy judgments of certain segmentals (Saito, 2013; Saito & Lyster, 2012 for English /r/; Lee & Lyster, 2017 for tense-lax vowel distinction [English /i/-/ɪ/]) and suprasegmentals (Bosker, Pinget, Quené, Sanders, & De Jong, 2013 for speed, breakdown and repair fluency; Parlak & Ziegler, 2017 for lexical stress; Munro & Derwing, 1995 for intonation). Given that what listeners can hear essentially matters for many researchers and practitioners alike, this expert rating approach has been most often used in primary pronunciation teaching studies as a way to assess the minimum, meaningful and perceptible units of L2 pronunciation (Thomson & Derwing, 2015).

Acoustic Analysis of Specific Proficiency

Certain researchers have further examined how instruction can affect pronunciation at a fine-grained level via computerized acoustic analyses of spectral (first, second and third formants of energy concentration [F1, F2, F3]), melodic (height and contour of fundamental frequency [F0]) and rhythmic/temporal (timing of sound and silence) information. Different from raters' segmental and suprasegmental judgements, the acoustic information is considered relatively objective in that it is automatically retrieved through speech analysis software, such as *Praat* (Boersma & Weenink, 2017), and presented at unambiguous and pre-determined units (hertz, milliseconds).[1] In order to interpret the results of acoustic analyses, the acoustic values need to be carefully cross-referenced with their relationship and impact on human perception (e.g., how listeners use different kinds of acoustic information to perceive the nativelike, accurate and intelligible form of specific segmental and suprasegmental sounds).

The use of these acoustic measures can be useful for researchers who are interested in the effect of instruction on different stages of L2 speech learning. For example, Lambacher, Martens, Kakehi, Marasinghe, and Molholt (2005) investigated whether instruction can stimulate L2 learners' abilities to distinguish the durational aspect of tense-lax vowel contrasts and change articulatory configurations, separately. From a theoretical perspective, these two abilities are hypothesized to correspond to the initial and later stages of L2 vowel acquisition, as learners seem to quickly acquire durational differences between L1 and L2 but may need much experience to decode spectral differences and relevant articulations (Bohn & Flege, 1997). In the context of fluency development, whereas between-clause pause ratio is considered to be relevant to L2 learners' conceptualization processes (i.e., figuring out what to say), within-clause pause ratio is believed to reflect their linguistic formulation processes (i.e., searching how to say in a target language) (Lambert, Kormos, & Minn, 2017). Using two different acoustic measures (between-clause vs. within-clause pause ratio) can provide a detailed picture of how instruction can help enhance L2 fluency as L2 learners attain more efficient, prompt and smooth conceptualization (i.e., reduction in between-clause pauses) and/or linguistic encoding (i.e., reduction in within-clause pauses).

At the same time, it is important to remember that acoustic information alone cannot fully simulate human speech perceptions. This is because human raters take into account not only speech signals for efficient and prompt word recognition, but a range of contextual factors as well (Broersma & Cutler, 2008). For example, only 50-60% of the variance in native listeners' perceptual judgments of nonnative speakers' English /r/-/l/ production can be explained by acoustic properties (i.e., degree and rate of F1, F2 and F3 information) (Flege, Takagi, & Mann, 1995; Saito & van Poeteren, 2018). This in turn supports the significant role of phonetic, lexical and contextual factors (Broersma & Cultler, 2008) and raters' particular accent familiarity (Bradlow & Bent, 2008) in L2 speech perception. To assess L2 learners' segmental and suprasegmental proficiency, therefore, researchers are strongly recommended

[1] Acoustic analyses cannot be fully objective, as they still entail some room for subjectivity. For example, researchers need to decide which acoustic parameters to highlight, where/when to locate a cursor and how to interpret the values in hertz/milliseconds. In this sense, acoustic analyses can serve as a *relatively* objective index of how pronunciation teaching can induce learners to acquire new manners and places of articulation when producing target sounds. If choosing to adopt these, researchers are required to carefully explain and justify which and how many acoustic dimensions they have chosen to focus on. For the acoustic analysis of L2 vowels, for example, a large number of parameters need to be included so as to reflect the dynamic, complex nature of the target features. In fact, many studies have carefully measured the F1, F2 and F3 dimensions of each sound at various time points of the sound's articulation (onset, 25%, 50% 75%, endpoint) (e.g., Oh, Guion-Anderson, Aoyama, Flege, Akahan-Yamada, & Yamada, 2011). In this way, researchers can, for example, capture how tongue movements take place while one vocalic sound is pronounced over time.

to justify their methodological decision to use acoustic analysis, human judgements or both in a complementary fashion according to the objectives of their studies.

To summarize this conceptual discussion on specific L2 pronunciation proficiency as per scoring methods (subjective vs. objective), the relationship between expert ratings and acoustic properties is summarized in Figure 3.

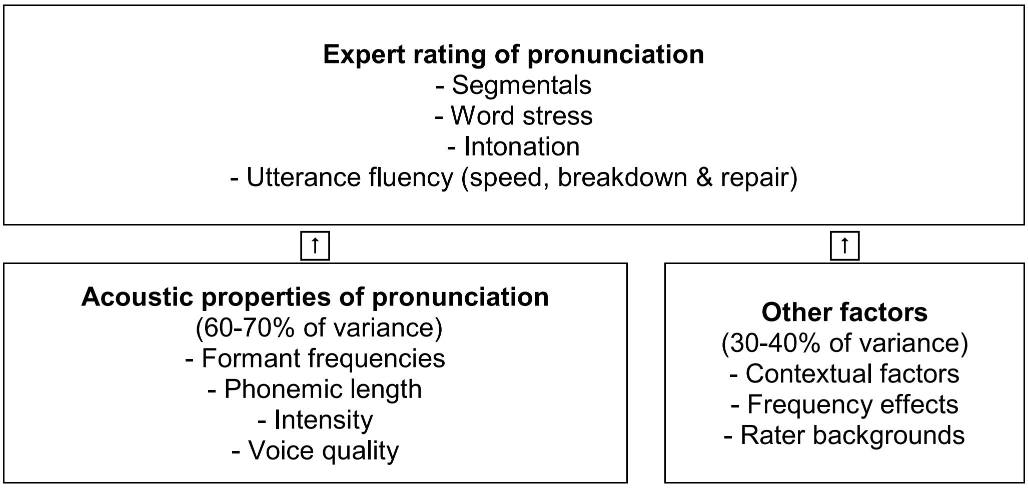

Figure 3. A Visual Summary of Specific L2 Pronunciation Proficiency as per Scoring Methods

Different Types of L2 Pronunciation Instruction and Resulting Knowledge

During pronunciation instruction, teachers explicitly provide metalinguistic information about target pronunciation features. As Derwing and Munro (2005) pointed out, "just as students learning certain grammar points benefit from being explicitly instructed…students learning L2 pronunciation benefit from being explicitly taught phonological form" (pp. 387-388). In grammar instruction, metalinguistic information generally refers to rule explanation of target morphosyntactic structures (e.g., plurality, tense, article). Looking at a range of intervention studies in L2 pronunciation teaching research, the nature of phonetic and phonological information that scholars provide during explicit instruction widely varies between primary studies. For example, many scholars have adopted articulatory-based instruction, where teachers help students understand the manner and place of articulation of L2 consonants and vowels relative to that of L1 counterparts. Such objectives can be achieved by using visual materials (e.g., diagrams, animations) (Celce-Murcia, Brinton, Goodwin, & Griner, 2010) and ultrasound imaging techniques (Gick, Bernhardt, Bacsfalvi, & Wilson, 2008). The notion of articulatory training echoes the gestural theory of L2 speech learning (e.g., Best & Tyler, 2007 for the direct-realist position). In this view, L2 speech information is thought to be stored in the brain based on relevant articulatory gestures (i.e., how to use the tongue, lips and jaw to produce new sounds). Such articulatorbased representations guide L2 learners to develop both perceptive and productive skills simultaneously.

On the other hand, other scholars have explored the efficacy of auditory-based instruction, where teachers help students grasp the perceptual dissimilarities and similarities between L2 sounds and L1 counterparts. For example, L2 leaners engage in intensive exposure to target features in different phonetic and lexical contexts, produced by multiple talkers (i.e., High Variability Phonetic Training) (Barriuso & Hayes-Harb, 2018). L2 learners also listen to acoustically enhanced input, either manipulated by signal processing techniques (e.g., Iverson, Hazan, & Bannister, 2005) or exaggerated by teachers (e.g., Saito, 2013). Such multivariate and hyper-articulated input is known to play a key role in both first and second language acquisition (Kuhl, 2004). The auditory training draws on a major theory of perception-based L2 speech learning: Learners store L2 speech information in the brain as per relevant spectral and temporal cues (how much they hear perceptual dissimilarities between new sounds and L1 counterparts) (Flege, 2016 for Speech Learning Model). Following this line of thought, any learning is hypothesized to initially happen in the dimension of perception skills, which will in turn activate relevant articulators (perception precedes production) (for a comprehensive overview on different theories in L2 speech learning, see Saito, 2018a).

Some attempts have been made to compare the differential effects of articulatory and auditory training on L2 perception and production development (e.g., Sakai, 2016). Importantly, many scholars have recommended using both options within the same syllabus, as there is a theoretical and empirical consensus that both perception and production dimensions are inter-related; and that stimulating both dimensions is believed to impact on L2 speech learning in a complementary fashion (Nagle, 2018). A growing number of scholars have also explored how to measure L2 learners' noticing, understanding and integration of metalinguistic information about the articulatory and auditory aspects of L2 sounds (for the roles and measurements of awareness in L2 speech learning, see Saito, 2018b; Venkatagiri & Levis, 2007). At the same time, such awareness factors may not be always a necessary condition for all types of successful L2 pronunciation learning, as some learning could take place without any awareness in naturalistic environments (e.g., Suzuki & DeKeyser, 2017).

In terms of classroom L2 pronunciation learning (the focus of the study), however, language is taught explicitly, and what teachers and students discuss is metalinguistic in nature, whether it is gesture-based (how to use articulators) or perception-based (how to hear new sounds). Despite these differences in concepts (articulatory vs. auditory), there are sufficient commonalities across all the studies (in that each of them encourages students to practice target phonological features with an aid of metalinguistic information) for us to group them as one instructional approach in the current study—i.e., explicit pronunciation instruction. In such contexts, what has actually remained open to debate concerns whether explicit instruction merely helps L2 learners accumulate metalinguistic information, or ultimately leads to acquisition and changes in performance at both controlled and spontaneous processing levels.

Turning to L2 grammar teaching research, scholars have thoroughly examined the robustness, transferability and durability of explicit instruction. For example, one well-researched topic has been what type of gain derives from instruction, with much discussion being directed towards conceptualizing, elaborating and refining the assessment frameworks for controlled and automatized L2 processing abilities, hereafter termed as *knowledge*. When Norris and Ortega (2000) conducted their first meta-analytic review of L2 grammar teaching studies published between 1980 and 1998, the effectiveness of instruction was identified as having a medium-to-large effect size. However, as the meta-analysts observed, these results drew on many primary studies that had used tasks that likely required more controlled, declarative knowledge (e.g., fill-in-the-blank tests) than more spontaneous, procedural knowledge (e.g., picture description, oral interview) (Doughty, 2003). Including more recently published L2 grammar teaching studies, Spada and Tomita (2010) conducted another meta-analysis, showing that different types of instruction help learners possess different types of L2 grammar knowledge. According to the results, the effectiveness of instruction is particularly large for relatively difficult, complex (as operationalized by the primary studies' authors) L2 features, even when measured via spontaneous tasks (for similar results of the effectiveness of one specific pedagogic technique—corrective feedback, see Lyster & Saito, 2010).

In this study, we argue that controlled and spontaneous L2 pronunciation performance should be considered as comprising different types of L2 knowledge which need to be assessed separately. Skill acquisition proponents (e.g., DeKeyser, 2017) have claimed that *instructed* SLA can be characterized as a gradual proceduralization and automatization of declarative knowledge. The metalinguistic information that students receive about the target language yields declarative knowledge (knowledge of facts and events). To build thorough declarative knowledge and integrate it into long-term memory, students first engage in controlled, form-focused exercises, where students practice the same target structures repetitively (e.g., fill-in-blanks, sentence combining). Drawing on such refined declarative knowledge as a main resource, students are then encouraged to create relevant procedural knowledge (knowledge about how to execute actions). This goal can be achieved by practicing the target structures when learers are required to prioritize meaning conveyance over linguistic accuracy (e.g., completing information gap tasks while using tense markers accurately)—the mid stage of skill acquisition referred to as "proceduralization."[2] As students gain an extensive amount of fully meaning-oriented L2 experience (e.g., immersion classrooms, study abroad), their procedural knowledge could be further restructured multiple times, and thus fine-tuned so that they can access the knowledge more quickly and accurately (and, it is usually thought, without or with less conscious control)—the final stage of skill acquisition referred as "automatization."

Following these conceptual discussions in instructed SLA more broadly, in the current study we use the term "controlled pronunciation knowledge" to refer to what L2 learners

[2] Certain scholars may disagree with the interface between declarative and procedural knowledge, arguably because much of procedural knowledge can be implicitly learned (e.g., Paradis, 2009). In this paper, however, we follow DeKeyser's (2017) position that whereas declarative and procedural knowledge are stored separately in the brain (i.e., temporal vs. frontal cortex), the activation of declarative knowledge through certain actions can stimulate the development of procedural knowledge simultaneously. After much of the same repetition, such procedural knowledge can be automatized, but may not necessarily become implicit (without any awareness) except certain individuals with unique cognitive profiles (e.g., implicit learning aptitude).

have learned from explicit phonetic instruction (e.g., articulatory [manner and place] and perceptual [duration, pitch and intensity] characteristics of target sounds). We use this term synonymously with declarative knowledge. "Spontaneous pronunciation knowledge" corresponds to the degree of L2 learners' access to explicit pronunciation knowledge in a timely manner while their primary focus is on meaning rather than form. We use this term synonymously with proceduralized and automatized knowledge. In the context of pronunciation teaching research, examining controlled speech performance is assumed to reveal the effect of instruction in the initial stages of L2 speech learning (i.e., noticing and consolidation of declarative knowledge); and examining spontaneous speech performance is assumed to reveal the role of instruction in the mid and later stages of L2 speech learning (i.e., proceduralization and automatization).

In the L2 pronunciation literature, certain scholars have similarly addressed the acquisition of different levels of knowledge in relation to increased experience and proficiency. For example, Major's (2001) Ontogeny Phylogeny Model describes L2 pronunciation development as a transition from L1-specific to universal errors. At the onset of L2 learning, L2 learners tend to substitute their own L1 counterparts for new L2 sounds. With increasing awareness of the new L2 sounds, interlanguage pronunciation development starts to take place where the learners attempt to use some composite forms between L1 and L2. At this stage, L2 learners' pronunciation forms likely reveal some universal characteristics, regardless of a learner's L1 background, such as a great deal of variation according to speech styles including different types of tasks. Namely, L2 learners' pronunciation forms tend to be more targetlike when their performance is elicited from formal controlled tasks (word reading) than from free speech tasks.

Focusing on Chinese learners of English in classroom settings, Rau, Chang, and Tarone (2009) showed that /θ/ was mispronounced more frequently in a picture description task than in word and sentence reading tasks. Although inexperienced Japanese learners' English /r/ pronunciation proficiency (length of residence < 1 year) was comparable across different task conditions (Saito & Munro, 2014), more experienced Japanese learners (length of residence > 1 year) produced English /r/ more accurately when their speech was elicited from word and sentence reading tasks than from picture description tasks (Saito & Brajot, 2013). This task effect in performance (controlled > spontaneous) was felt to arise from the increased demands on linguistic processing due to the lack of substantial planning time for the picture description tasks, compared to the more controlled reading tasks (for similar results of consonant cluster, see also Lin, 2003).

Although the Ontogeny Phylogeny Model and its relevant findings reviewed here do not directly concern pronunciation teaching, it has some implications. Since, according to that model, L2 learners' speech generally develops at different learning rates (controlled → spontaneous knowledge), it is reasonable to assume that similar developmental patterns could be observed when L2 learners receive some form of explicit instruction. Crucially, the Ontogeny Phylogeny Model stresses that task performance variation serves as universal evidence of interlanguage development (away from L1 substitutions), and slowly declines as a function of increased L2 experience and proficiency. Therefore, a range of controlled and spontaneous tasks need to be adopted in order to adequately assess the extent to which instructional gains can be transferrable when learners use different types of speaking styles in diverse speaking contexts (more formal and controlled vs. more free and spontaneous).

As operationalized in Spada and Tomita's (2010) coding scheme, controlled speech tasks are designed to elicit more explicit, analyzed and conscious knowledge of L2 pronunciation forms, i.e., when learners can fully monitor their accurate use of target features by accessing newly-learned metalinguistic knowledge. In the context of L2 pronunciation teaching, for example, Couper (2006) asked participants to read sentences which included words with different types of consonant clusters in different positions (e.g., *climber, difficult*) (i.e., a sentence reading task). In Munro and Derwing (2008), participants repeated audio prompts ("*the next word is* _____") where different target words which included various L2 English vowels (e.g., *beat, bit, bet, bait*) were inserted in the blank (i.e., a delayed repetition task).

In contrast with controlled tasks, spontaneous speech tasks are better suited to measure L2 learners' relatively unconscious and unmonitored use of L2 pronunciation forms. In these tasks, L2 learners are guided to pay simultaneous attention to the grammatical, phonological, lexical, and pragmatic aspects of language to convey their intended message as a primary focus; the tasks are executed within a realistic time limit so that L2 learners will not have much planning time to access explicit articulatory knowledge. For instance, Parlak and Ziegler (2017) asked participants to read given sentences including a range of polysyllabic target words (e.g., *the most imPORtant comPOnent of a good LANguage lesson is*…), and share/exchange their opinions accordingly. The extent to which they accurately placed word stress while pronouncing the target words was analyzed acoustically. In another example, Saito and Brajot (2013) asked participants to describe a series of pictures with five seconds of planning time (i.e., timed picture description). Each picture had three key words that they had to use in their descriptions, one of which was a target word including English /r/ (e.g., *read, road*).

It is important to point out that the spontaneous tasks described here can be somewhat semi-structured (rather than entirely free), as they are intended to elicit participants' use of specific phonological features; thus, they are essentially different from unstructured, extemporaneous tasks (e.g., cartoon description, oral interview), where participants freely use language to describe a sequence of events or answer questions without any written/oral prompts. Given that outcome measures in intervention studies are supposed to reflect the focus of instruction (Ortega, 2003), we consider it crucial to use spontaneous/semi-structured tasks as well as extemporaneous and unstructured tasks[3] to precisely examine how much L2 learners improve on what they have been taught. The effectiveness of instruction needs to be assessed, not only when they are allowed to solely focus on accurate pronunciation forms

[3] It is important that intervention studies should also look at transferability to other kinds of tasks, where participants might not expect to use the feature they were taught (i.e., extemporaneous and unstructured tasks). While some studies have taken an exploratory approach towards examining how L2 speakers pronounce particular features during entirely free speaking tasks, analyses need to be conducted with much caution as pronunciation forms can be subject to phonetic and lexical contexts and the number of observations may vary between participants (see Piske, Flege, MacKay, & Meador, 2011).

(i.e., controlled knowledge), but also when their primary focus lies in using the language to communicate (i.e., spontaneous knowledge).

A final point we would like to make is that our admittedly simplified framework of controlled and spontaneous knowledge reflects what is crucial for teachers and learners and what instructed SLA research is mainly concerned with—i.e., the extent to which L2 learners have automatized explicit/controlled knowledge resulting from instruction (proceduralization and automatization). In keeping with the instructed SLA paradigm (e.g., DeKeyser, 2017; Spada & Tomita, 2010), our assumption is that this rather "rough" distinction can be measured via controlled and spontaneous tasks. In the former task format, L2 learners are allowed to solely focus on pronunciation accuracy (e.g., word, sentence and paragraph reading tasks); and in the latter task format, their L2 pronunciation accuracy is tested when they primarily use language for meaning (e.g., picture description, naming and narrative). In essence, their task performance can be thought to mirror their controlled and spontaneous knowledge, a distinction which is important especially in classroom settings where the ultimate goal of both teachers and students is to ensure that what is taught explicitly can be transferred outside the classroom (Gatbonton & Segalowitz, 2005). For more theoretically important dimensions of task-knowledge interaction (e.g., explicit vs. implicit), see our arguments in the Discussion section. The proposed construct of L2 pronunciation knowledge is summarized in Table 1.

Table 1 *A Summary of Different Types of L2 Pronunciation Knowledge and Corresponding Elicitation Tasks*

Target knowledge	Task type	Corresponding skill acquisition stage	Task features
Controlled knowledge	Controlled speech task	Consolidating declarative knowledge	- structured - focus only on linguistic accuracy - no time pressure
Spontaneous knowledge	Spontaneous task	Proceduralization and automatization	- semi-structured - 1st focus on meaning and 2nd focus on form - time pressure

Research Synthesis and Meta-Analysis

In this paper, we argue that any impact of instruction on L2 pronunciation needs to be separately examined according to measurement focus, scoring method and task type, as they tap into different dimensions of L2 pronunciation proficiency and knowledge. To this end, we proposed our model on how to measure and interpret the efficacy of pronunciation teaching—i.e., the Framework for L2 Pronunciation Measurement. To provide a new look at the current state of affairs in L2 pronunciation research, and test the validity of our arguments, the second objective of this current project is to survey measurement practices in instructed L2 pronunciation research and re-examine the effectiveness of pronunciation teaching through our proposed assessment framework.

To achieve these goals, we first retrieved a dataset of 77 primary studies published between 1982 and 2017. The measurement practices found in this sample were then coded and synthesized vis-à-vis focus of measurement (global vs. specific), scoring methods (subjective vs. objective) and knowledge/task types (controlled vs. spontaneous). As we addressed earlier, Levis's (2005) special issue in *TESOL Quarterly* could be considered as a turning point for L2 pronunciation teaching research. The issue featured a set of influential position papers which identified the marginalization of L2 pronunciation research in applied linguistics, made a strong call for future research, and provided a range of suggestions for methodological rigor (e.g., Derwing & Munro, 2005). Whereas our dataset covers a range of publications over an extensive period of time (1982-2017), we assume that the quality of primary research has changed to a great degree since Levis (2005). Given that scholars likely spend a few years designing, implementing and publishing their work, we survey the extent to which different types of L2 pronunciation measurements (as classified by our proposed framework) have been used in studies of instructed L2 pronunciation in accordance with two different timeframes—relatively traditional studies published before the proposed paradigm shift (up until 2007: 1982-2007) (n = 22) and more recent studies published over the past 10 years (from 2008: 2008-2017) (n = 55). Three research questions are formulated:

1. How often have global and specific L2 pronunciation proficiency been analyzed in primary studies up until 2007 and from 2008?
2. How often have expert judgements and acoustic analyses been employed in primary studies until 2007 and from 2008?
3. How often have controlled and spontaneous tasks been incorporated in primary studies until 2007 and from 2008?

As part of a growing line of research in the 'SLA-assessment interface' (see e.g., Bachman, 1988; Bachman & Palmer, 1998; Derrick 2016; Norris & Ortega, 2012), we are also interested in the psychometric properties of different pronunciation measurements. Toward this end, we examine variability in inter-rater reliability estimates. Syntheses of this nature, referred to as 'reliability generalization meta-analysis', aggregate reliability coefficients in order to (a) approximate population-level estimates of reliability and/or to (b) examine systematic variability in observed estimates across study, sample, or measurement related features (Wheeler, Vassar, Worley, & Barnes, 2011). One recent example of this perhaps less familiar meta-analytic approach in the context of instructed SLA can be found in Plonsky and Derrick's (2016) study. The authors were interested in estimating the distribution of reliability coefficients as a means to aid researchers' attempts to interpret them. To this end, 2,244 reliability estimates were extracted from 537 primary studies along with a number of sample, design, and measurement-related features. Among other findings, the results of their analysis revealed the median estimate for interrater reliability of L2 pronunciation to be .81 (IQR = .2, based on *k* = 55). We are interested in further understanding whether a similar standard would be generalizable to the measurements in instructed L2 pronunciation research. The next research question, therefore, is:

4. What is the observed reliability (interrater) of outcome measures used in L2 pronunciation instruction overall and across different types of assessments?

Finally, by taking a meta-analytic approach, we elucidate the effects of pronunciation teaching in a representative sample of primary studies according to our proposed framework. As such, the findings of this project are expected to allow us to probe the extent to which pronunciation teaching can impact L2 learners' specific pronunciation forms at perceptible and fine-grained levels and then relate to listeners' global impression of overall pronunciation (i.e., instruction and proficiency link); and the extent to which pronunciation teaching can lead to substantial change in L2 learners' controlled production at the initial stage of L2 speech learning (noticing, error avoidance) as well as their spontaneous production at the later stage of L2 speech learning (automatization) (i.e., pronunciation teaching and knowledge link). As operationalized in virtually all meta-analyses of instructed SLA (e.g., Norris & Ortega, 2000), the magnitude of pronunciation teaching gain is calculated via Cohen's *d*, and interpreted from "small" to "large" as per Plonsky and Oswald's (2014) benchmark. The final research question is formulated as follows:

5. To what extent do the effects of L2 pronunciation instruction vary across different types of measurement focus (global vs. specific), scoring method (subjective vs. objective) and task type (controlled vs. spontaneous)?

Study Retrieval

Following the search/selection procedure described in the precursor meta-analysis study (i.e., Lee et al., 2015), we aimed to identify primary studies which examined learners' gain with a pre-posttest design (within groups), when they received instruction on one or more pronunciation features, or when their performance was compared to that of a comparison group receiving no pronunciation instruction (between groups). As in Lee et al. (2015), our analyses focused on studies which examined either students' improvement over time (within groups) or the comparison between experimental and control groups as well as those which featured both within and between group analyses. Using combinations of key words (second language, foreign language, pronunciation, instruction), we searched major library databases (e.g., Educational Resources Information Center, Linguistics and Language Behavior Abstracts, PsycINFO, PsycArticles, Web of Science) and online resources (e.g., Google, Google Scholar).

At the same time, the scope of our research synthesis and meta-analysis departed from Lee et al. (2015) in the following respects. First, whereas Lee et al.'s search concluded with studies published in 2013, ours extended to December 2017. As such, we aimed to reveal the current state of knowledge in the field of pronunciation teaching research. This decision was crucial, as the number of pronunciation teaching studies has continued to grow over the past four years (2014-2017) to a great degree (see below).

Second, we narrowed down our search, focusing exclusively on published journal articles without including so-called fugitive literature (e.g., unpublished doctoral dissertations, conference proceedings and presentations, or book chapters). This decision was substantially different from Lee et al.'s (2015) meta-analysis which took an inclusive approach (featuring journal articles, book chapters, and conference proceedings/presentations). We acknowledge that our approach (i.e., focusing on published journal articles) may entail a degree of publication bias since statistically significant results are more often reported and accepted for publication (see Plonsky & Oswald, 2014).

It needs to be stressed that the main focus of the current project lied in the examination of the publication practices especially after Levis's (2005) special issue in *TESOL Quarterly*, which we consider as the landmark publication for the development of L2 pronunciation research. In the same issue, Derwing and Munro (2005) posited that "the study of pronunciation has been marginalized within the field of applied linguistics" (p. 379) while providing the results of their brief survey on the number of published papers in major academic journals.

In response to Derwing and Munro (2005), our project looked at the current state of knowledge in the context of published articles all of which went through thorough peer review. In so doing, our goal was to compare the publication trends over the past 10 years (2008-2017) compared to the marginalization observed in 2005; and to examine the relative status of pronunciation teaching research within a broader framework of instructed SLA research. Ultimately, we aimed to reveal the extent to which L2 pronunciation research has grown and how its growth aligns with methodological rigor in accessible, peer-reviewed literature most likely to have an influence on SLA theory and practice (see rationales for the same decision in Marsden, Morgan-Short, Thompson, & Abugaber, 2018; Marsden, Thompson & Plonsky, 2018; Norris & Ortega, 2000; Plonsky, Marsden, Crowther, Gass, & Spinner, in press; Spada & Tomita, 2010).

Dataset

The final dataset comprised 77 intervention studies, which examined the effect of instruction on L2 pronunciation proficiency adopting a pre-and-post-test design, published between 1982 and 2017. Among 77 individual studies, 47 were included in Lee et al. (2015); and 30 were newly added (see the studies listed in **Supporting Information-A**). In line with Derwing and Munro's (2005) observation of the availability of "relatively little published research on pronunciation teaching" (p. 383), only 22 articles were published in peer review journals between 1982 and 2007. Comparatively, a total of 55 studies have been conceptualized, conducted and published over the past 10 years (2008-2017), a positive uplift possibly triggered by Levis's (2005) special issue. We also note that the dataset (*N* = 77 primary studies between 1982 and 2017) is substantially greater than most other syntheses and meta-analyses in instructed SLA research (e.g., Goo et al., 2015), allowing for fairly robust conclusions.

In total, there were 2,573 participants including 1,961 participants who received pronunciation teaching and 612 who did not (i.e., control participants). In our entire dataset, 67 studies examined within-group contrasts (comparing pre- and post-test performance) and 39 studies probed between-group contrasts (comparing post-test performance of experimental and control participants). Twenty-nine studies featured the analyses of both within- and between-group comparisons. Thirty-eight studies focused on only within-group contrasts[4]; and 10 studies focused only on between-group contrasts. The sample of primary studies was coded for the three crucial elements of L2 pronunciation assessment highlighted in the proposed framework: (a) focus of measurement, (b) scoring method and (c) task type, defined as follows.

Focus of Measurement. This variable corresponds to which aspects of L2 pronunciation proficiency (specific vs. global) primary studies aimed to test. Many studies concerned L2 learners' acquisition of specific segmental (consonants, vowels) (e.g., Saito, 2013), syllabic (schwa vowel insertion) (e.g., Couper, 2006), prosodic (word and sentence stress, intonation) (e.g., Kennedy, Blanchet, & Trofimovich, 2014), and temporal features (speed, breakdown, repair fluency) (e.g., De Jong & Perfetti, 2011). We include both prosodic and temporal features as specific constructs of L2 pronunciation proficiency, as they are considered to belong to the same category of suprasegmentals (for details of the distinction between the melody- and rhythm-based suprasegmental features, see Trofimovich & Baker, 2006). Some scholars also examined the impact of pronunciation teaching on global constructs of L2 pronunciation proficiency (i.e., comprehensibility, intelligibility, perceived fluency, accentedness) (e.g., Derwing et al., 2014).

Scoring Method. This variable corresponds to how researchers analyzed the impact of pronunciation teaching. Three subcategories were coded for. First, global proficiency was operationalized via listeners' intuitive judgements of comprehensibility, perceived fluency and accentedness (e.g., Levis, Sonsaat, Link, & Barriuso, 2016). Under this, intelligibility was analyzed not only through listeners' transcription (e.g., Derwing et al., 2014), but also through their judgements (e.g., Martinsen, Montgomery, & Willardson, 2017). Second, experts' ratings of specific pronunciation proficiency were operationalized via experts' ratings of segmental (e.g., Lee & Lyster, 2017), prosodic (e.g., Hardison, 2005) and temporal (e.g., Gorsuch, 2011) qualities. Third, specific pronunciation proficiency was operationalized

[4] It is worth mentioning that adopting only within-group contrasts is known to be weaker; effect sizes from such designs should be treated with caution, because validity could be threatened by regression to the mean, test effect, maturation and other factors).

via acoustic analyses of segmental (e.g., Offerman & Olson, 2016), prosodic (e.g., Parlak & Ziegler, 2017) and temporal (e.g., De Jong & Perfetti, 2011) information.

Task Type. This variable corresponds to what types of outcome measures were used in primary pronunciation teaching studies. Among a range of moderator analyses, Lee et al.'s (2015) meta-analysis took a first step towards examining the relative effectiveness of pronunciation instruction according to two broad categories of outcome measures—more controlled vs. more spontaneous tasks. According to the results, L2 learners' gain was significantly larger when their production was elicited from the controlled than the spontaneous tasks (*d* = 0.96, [.89, 1.00] vs. 0.37 [.30, .44] for between-group, *d* = 0.96, [.90, 1.02] vs. 0.65 [.59, .71] for within-group). However, Lee et al. did not explain precisely how they coded controlled and spontaneous measures. This is arguably because it is extremely difficult to conduct such moderator analyses without proposing, establishing and justifying a strong measurement model of this kind. In the current investigation, we extended their preliminary investigation in the following manner.

First, we looked at the role of task type at a more fine-grained level by separately calculating effect sizes in terms of focus of measurement (global vs. specific) and scoring methods (subjective vs. objective). Second, we focused only on studies published in peer-reviewed journals. Lee et al. (2015) included a range of publications (including conference proceedings and presentations) but noted that much detailed information remained unreported despite the fact that measuring spontaneous L2 pronunciation requires much caution and careful analyses (Piske et al., 2011). Third, we included more recent studies (*n* = 30 studies published between 2014 and 2017). Adding the newer dataset is crucial, as the field of pronunciation teaching research has witnessed a great deal of methodological discussion, renovation and innovation over the past few years. For example, Saito's (2012) research synthesis identified the exclusive reliance on controlled tasks (which can better reflect how much learners can monitor their correct pronunciation inside class) while making a strong call for adopting and elaborating spontaneous tasks (which can better index how much learners can use learned knowledge outside classrooms).

Following Spada and Tomita's task taxonomy, we first featured controlled production tasks, wherein L2 learners were allowed to focus *solely* on accurate and fluent use of language, such as word, sentence and paragraph reading (Saito & Saito, 2017) and delayed repetition tasks (Lee & Lyster, 2017). Out second coded category was spontaneous production tasks, wherein L2 learners were guided to use language accurately and fluently while at the same time using language for meaning as a primary focus, such as picture naming (Offerman & Olson, 2016), picture narrative (Trofimovich et al., 2017), timed picture description (Saito, 2013), and interview (Parlak & Ziegler, 2016).

Effect Size Analyses

For all the between- (experimental vs. control) and within-group (pre vs. post-tests) contrasts, Cohen's *d* index was calculated for between- and within-group contrasts by using the pooled between-groups standard deviation:

$$d = \frac{M_1 - M_2}{\sigma_{\text{pooled}}}$$

$$\sigma_{\text{pooled}} = \frac{(n_1 - 1)\,\sigma_1 + (n_2 - 1)\,\sigma_2}{(n_1 - 1) + (n_2 - 1)}$$

where *M* is the mean and σ is the standard deviation,

Finally, the mean effect sizes were calculated and weighted to the sample size of primary studies according to six subcategories: (1) Global/Controlled, (2) Global/Spontaneous, (3) Specific/Subjective/Controlled, (4) Specific/Subjective/Spontaneous, (5) Specific/Objective/Controlled and (6) Specific/Objective/Spontaneous. A total of 119 effect sizes (Cohen's *d*) were calculated for within-group contrasts; and 68 effect sizes based on between-group contrasts.[5]

Building on Plonsky and Oswald's (2014) procedure for inter-coding of meta-analytic data, the first and second authors first independently coded *n* = 77 pronunciation teaching studies (see **Supporting Information-A**). Then, we compared/discussed results to ensure that our understanding of the coding scheme was consistent according to the proposed framework. At this stage, our agreement rate was 96.1% (74 out of 77 studies). Afterwards, both of the authors engaged in discussion and adjustment together, when any disagreement emerged and/or where clarification was needed to come to a common decision (see our final results in **Supporting Information-B**).

Reliability Estimates

Finally, we extracted reliability estimates when available. In the current analysis, we are particularly concerned with the extent which the effectiveness of instruction could vary when listeners and raters analyze global (e.g., comprehensibility) and specific (e.g., segmental and suprasegmental accuracy) aspects of L2 pronunciation proficiency. Following our proposed framework, therefore, inter-rater reliability was retrieved and averaged from the

[5] For the purpose of comparability, we focused on the results of immediate post-tests in the current meta-analysis. However, we noted that a growing number of recent pronunciation teaching studies have adopted not only immediate, but also delayed post-tests. Future synthesis studies should pursue the durability of instruction by looking different intervals of post-tests as a potential moderator variable.

following four subcontexts: listeners' impressionistic judgements of global proficiency (comprehensibility, intelligibility, accentedness, fluency) and experts' ratings of specific proficiency (segmentals, suprasegmentals) under two task conditions (controlled, spontaneous): (1) Global/Controlled, (2) Global/Spontaneous, (3) Specific/Subjective/Controlled, and (4) Specific/Subjective/Spontaneous.

Analysis

In order to address RQs 1-3, concerning the use of different types of measures as laid out in our framework, frequencies and percentages of each were calculated. RQ 4 and RQ 5 were conducted in parallel as both addressed variability as a function of different measurement options. Both RQs sought to provide complementary angles on the different subcomponents of pronunciation teaching measurements in our model. More specifically, to address RQ 4, the different measures in our framework were used to group and meta-analyze reliability estimates in our sample. Similarly, for RQ 5, subgroups were formed based on the same measurement types; in this case, however, the focus of our analysis was on the effect sizes (and the corresponding descriptive statistics) associated with those groupings.

Results

Research Synthesis of Pronunciation Teaching Research Practices

The first aim of the results section is to describe the extent to which primary studies have measured the effectiveness of pronunciation instruction across the three key methodological dimensions: (a) focus of measurement (global vs. specific proficiency); (b) scoring method (expert ratings vs. acoustic analyses); and (c) task type (controlled vs. spontaneous tasks). In addition, we survey the extent to which such research practices have changed over the past 10 years (2008-2017).

Focus of Measurement. It is important to remember that all the primary studies in this current meta-analysis were concerned with helping students improve their specific pronunciation proficiency (segmental and suprasegmental accuracy) via explicit phonetic instruction. However, Table 2 demonstrates some inconsistency in terms of the focus of their outcome measures. The results of the syntheses demonstrated that a majority of pronunciation teaching research (54 out of 77 studies = 70.1%) focused on how explicit instruction (teaching segmental/suprasegmental accuracy) could promote the development of specific L2 pronunciation proficiency (segmental and suprasegmental accuracy). A total of 8 out of 77 studies (10.3%) investigated the extent to which such instruction could ultimately impact on L2 learners' global pronunciation proficiency as well (comprehensibility, intelligibility, perceived fluency, accentedness). Notably, 15 out of 77 studies (19.4%) adopted only global L2 pronunciation proficiency measures as a way to evaluate the instructional effectiveness (teaching segmental and suprasegmental accuracy), but without any mention of its impact on specific L2 pronunciation proficiency.

In terms of research trends between 1982 and 2017, the number of publications has doubled over the past 10 years (*n* = 22 in 1982-2007 → 55 in 2008-2017). Compared to studies up to 2007, however, studies after 2008 featured both global and specific measures slightly more frequently between 2008 and 2017 (9.0% → 10.9%). Surprisingly, more recent studies seem to be more likely to adopt only global measures (e.g., comprehensibility) than previously despite the fact that they taught specific pronunciation proficiency (segmental and suprasegmental accuracy), exhibiting more observable mismatch between the focus of instruction and measurement (13.6% → 23.6%).

Table 2 *Summary of Measurement Focus of N = 77 Primary Pronunciation Teaching Studies*

	Global only	Specific only	Both global and specific
A. Total (*n* = 77)			
No. of primary studies	15	54	8
Ratio (out of 77 studies)	19.4%	70.1%	10.3%
B. Up until 2007 (*n* = 22)			
No. of primary studies	3	17	2
Ratio (out of 22 studies)	13.6%	77.2%	9.0%
C. From 2008 (*n* = 55)			
No. of primary studies	12	37	6
Ratio (out of 55 studies)	23.6%	67.2%	10.9%

Scoring Method. In terms of scoring methods, out of 62 studies which examined the impact of instruction on specific L2 pronunciation proficiency, slightly more researchers appeared to prefer experts' ratings (39 out of 62 studies = 62.9%) to acoustic and objective analyses (34 out of 62 studies = 54.8%) in order to examine the impact of pronunciation teaching on segmental and suprasegmental accuracy (summarized in Table 3). A small portion of primary studies (11 out of 62 studies = 17.7%) adopted both subjective and objective measures. According to our sub-analysis (publications up until 2007 vs. from 2008), such positive changes seem to have been accelerated over the past 10 years. The number of primary studies which incorporated both expert ratings and acoustic analyses has substantially increased so as to better capture the instructional gains at perceptible and fine-grained levels (10.5% → 20.9%).

Table 3 *Summary of Scoring Method of N = 62 Primary Pronunciation Teaching Studies*

	Expert ratings only	Acoustic analysis only	Both expert ratings and acoustic analysis
A. Total (*n* = 62)			
No. of primary studies	28	23	11
Ratio (out of 61 studies)	45.1%	37.0%	17.7%
B. Up until 2007 (*n* = 19)			
No. of primary studies	11	6	2
Ratio (out of 19 studies)	57.8%	31.5%	10.5%
C. From 2008 (*n* = 43)			
No. of primary studies	17	17	9
Ratio (out of 42 studies)	39.5%	39.5%	20.9%

Task Type. As shown in Table 4, the results of the synthesis identified some evidence of imbalance in task type in the existing pronunciation teaching literature. To examine the acquisitional value of pronunciation teaching, 44 out of 77 studies adopted only controlled tasks (57.1%), and 14 studies employed only spontaneous tasks (18.1%). It was only 19 out of 77 studies (24.6%), where researchers made efforts to adopt both controlled and spontaneous speech tasks.

Interestingly, the imbalance between controlled and spontaneous tasks has been gradually reversed over the past 10 years. Comparing practices in publications up until 2007 and from 2008 (1982-2007 vs. 2008-2017), scholars relied less on controlled measures only (77.2 → 49.0%), and included more spontaneous measures (9.0% → 21.8%). A growing number of scholars have employed multiple measures to track the impact of instruction on both students' controlled and spontaneous L2 pronunciation proficiency (13.6% → 29.0%).

Table 4 *Summary of Task Type of N = 77 Primary Pronunciation Teaching Studies*

	Controlled only	Spontaneous only	Both controlled and spontaneous
A. Total (*n* = 77)			
No. of primary studies	44	14	19
Ratio (out of 77 studies)	57.1%	18.1%	24.6%
B. Up until 2007 (*n* = 22)			
No. of primary studies	17	2	3
Ratio (out of 22 studies)	77.2%	9.0%	13.6%
C. From 2008 (*n* = 55)			
No. of primary studies	27	12	16
Ratio (out of 55 studies)	49.0%	21.8%	29.0%

Reliability in Pronunciation Teaching Research

The second aim of the results section is to examine the degree of error in measures of the effectiveness of pronunciation teaching. Not surprisingly, authors in this domain have relied on a variety of different indices including kappa as well as we percent agreement, and Cronbach's alpha, depending on the scoring and data that were collected. They were calculated to illustrate the degree of agreement among the listeners and raters' subjective judgements. We would first point out that such estimates were available for 67.9% of our sample (36 out of 53 studies which adopted global or/and expert judgments as outcome measures). For research practices up until 2007 and from 2008, we found positive trends that more primary studies have begun to report reliability (50.0% → 75.6%) (see Table 5).

Table 5 *Summary of Reliability Report of N = 53 Primary Pronunciation Teaching Studies*

	Reported
A. Total (*n* = 53)	
No. of primary studies	36
Ratio (out of 53 studies)	67.9%
B. Up until 2007 (*n* = 16)	
No. of primary studies	8
Ratio (out of 16 studies)	50.0%
C. From 2008 (*n* = 37)	
No. of primary studies	28
Ratio (out of 37 studies)	75.6%

Next, we conducted the reliability generalization meta-analysis resulting in following observations (summarized in Table 6). We calculated reliability for the subjective analyses (listeners' intuitive assessments of global proficiency and expert coders' ratings of specific proficiency). We did not do so for the objective analyses, wherein all the analyses were computed via relevant software resulting in less room for subjectivity (e.g., *Praat*) and α was hypothesized to be 1.00. First, reliability estimates in the realm of pronunciation teaching are generally quite high ranging overall from .76 to .93. However, there are several patterns of variability worth noting. For example, reliability estimates tend to be higher for listeners' judgements of global proficiency (.93, .87) than for experts' ratings of specific linguistic targets (.86, .76). Furthermore, regardless of the focus of the measurements, there is less measurement error when learner production is controlled as opposed to spontaneous (.93 vs. .87; .86 vs. .76). It is also worth noting that the type of measure most often used (global proficiency with controlled production) is also the one with the highest reliability (α = .93).

Table 6

Median Reliability according to Measurement Focus, Scoring Methods and Task Types

Proficiency dimension	Global		Specific	
Scoring method	Human intuition		Expert coding	
Task/knowledge type	Controlled production	Spontaneous production	Controlled production	Spontaneous production
No. of primary studies	16	19	32	7
Ratio (out of 74)	21.6%	25.7%	43.2%	9.5%
Median Reliability	.93	.87	.86	.76
Interquartile range	0.16	0.17	0.15	0.13

Effect Size Analyses

The final aim of the results section is to illustrate the extent to which the effectiveness of instruction could vary according to three crucial dimensions of L2 pronunciation proficiency and knowledge—measurement focus, scoring method and task type. Below, we report mean effect sizes and their 95% CI values according to six subcategories (specified by our proposed model and summarized in Table 7). All the effect size indices were interpreted with reference to Plonsky and Oswald's (2014) benchmarks:

- For between-group contrasts, *d* = 0.4 for small, 0.7 for medium and 1.0 for large.
- For within-group contrasts, *d* = 0.6 for small, 1.0 for medium and 1.4 for large.

Publication Bias. In order to assess the presence of bias in the collected evidence toward statistically significant findings, we have produced two funnel plots: one for between- and another for within-group contrasts (see Figures 4 and 5). In the absence of bias (i.e., in a true, normal distribution of effects), we would expect to see fairly even dispersion on either side of the mean effect. A wider spread is also anticipated lower in the figure due to the enhanced variability in effects derived from smaller samples. The data for the present study reveal substantial variability up and down the plots. This is not entirely surprising considering that even the relatively 'larger' studies in our sample are based on fairly small *N*s. Such variability in effects can also be explained by the wide range of variables found to moderate the effects of pronunciation teaching including, as we show below, different outcome measures. It is somewhat concerning, however, that the spread of effects seems wider to the right of the means in these two figures. With only a few exceptions, in fact, the distribution of effects in the within-groups sample appears almost cut-off at zero. Such 'missing values', would seem to indicate a possible suppression of smaller and/or negative effects, relative to an unobstructed population of effects. Consequently, we might consider the overall effects from this study to provide a slightly inflated estimate of the effect of pronunciation teaching.

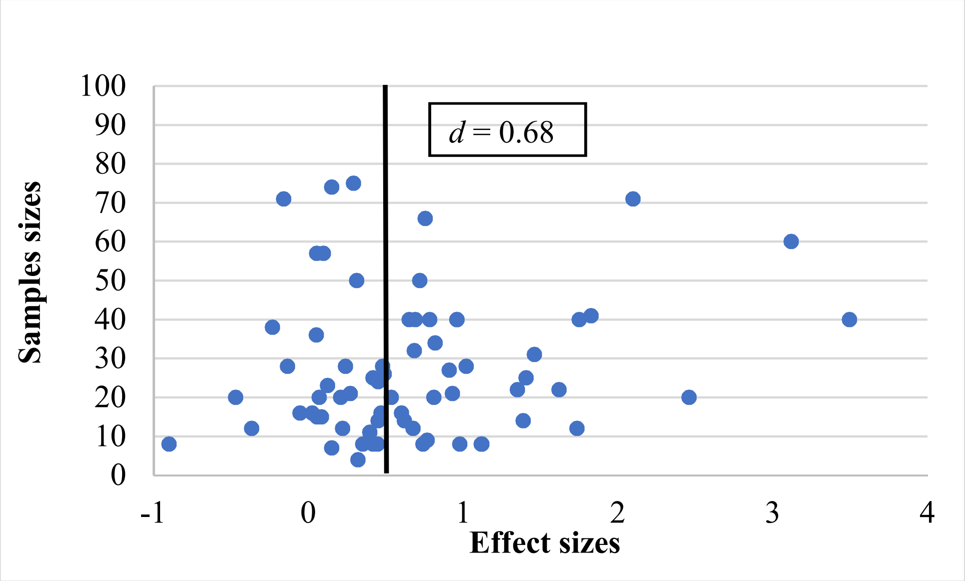

Figure 4. Funnel plot for between-groups effect sizes
Note. Effect sizes and sample sizes are plotted on the x- and y-axes, respectively. The average effect appears at the top of the figure.

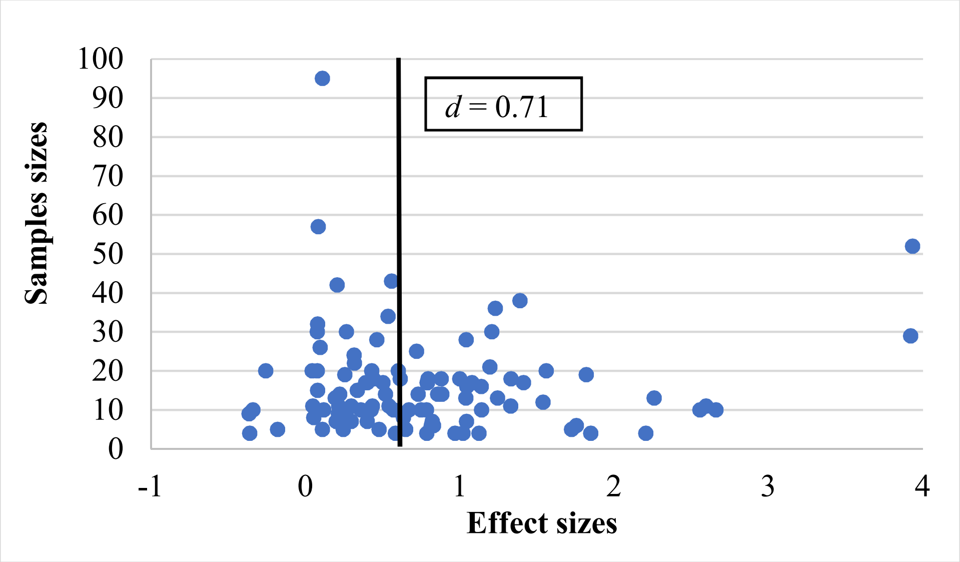

Figure 5. Funnel plot for between-groups effect sizes
Note. Effect sizes and sample sizes are plotted on the x- and y-axes, respectively. The average effect appears at the top of the figure.

Overall Effectiveness. Similar to the precursor meta-analysis (Lee e al., 2015), the results of the effect size analyses demonstrated that pronunciation teaching positively influenced L2 pronunciation development for between-group contrasts (*d* = 0.68, *95% CI* = 0.49-0.86) and for within-group contrasts (*d* = 0.73, *95% CI* = 0.69-0.78). For between-group comparisons, this overall effect size can be seen as both statistically reliable and stable (given their 95% CIs do not include zero). According to Plonsky and Oswald's (2014) framework of reference often adopted in L2 research, the 95% CI values here suggested that the size of pronunciation teaching effects could be considered roughly medium for between-group comparisons, relative to L2 research in general, and perhaps more in the small-to-medium range for the domain of instructed SLA (see Plonsky, 2017).

As for within-group comparisons, the average *d* value (0.73) could be considered as small for within-group comparisons (within the $0.6 < d < 1.0$ range of 'small effects' found by Plonsky & Oswald). It needs to be noted that the effect size of within-group comparisons may indicate not only the efficacy of instruction, but also test-retest effects or other maturational or exposure-related effects. To interpret of the results adequately, we calculated Cohen's *d* value for the pre- and post-test results of control groups in a total of 29 studies comprising 612 participants, $d = 0.31$, *95% CI* =0.24-0.38. The results indicated that the experimental groups receiving pronunciation instruction demonstrate significant improvement in their L2 pronunciation proficiency compared to control groups who do not as indicated by distance between the two groups' confidence intervals (*95% CI* = 0.69-0.78 vs. 0.24-0.38).

To re-examine the effectiveness of pronunciation teaching as per our proposed framework, all of the collected *d* values were analyzed in terms of measurement focus (global vs. specific), scoring method (subjective vs. objective), and task type (controlled vs. spontaneous). Table 7 summarizes this phase of the analysis, showing the differential effects of pronunciation teaching (*d* values) according to the different dimensions of L2 pronunciation proficiency and knowledge. For each dimension, weighted mean effect sizes and 95% CIs were calculated. For between-group comparisons (experimental vs. comparison at post-tests), the effects of instruction were considered significant when CIs did not cross zero. For within-group comparisons, experimental groups' instructional effectiveness (pre-post) needs to be detangled from test-retest effects. To this end, we calculated 95% *CI*s of comparison groups (who just took tests twice without any pronunciation instruction), *d* = 0.24-0.38. Experimental groups' gains were interpreted as significant when the *CIs* of the experimental groups went beyond the upper range of comparison groups' *CIs* ($d > 0.38$). Any comparison analyses were conducted as per focus of measurement (global vs. specific), scoring method (subjective vs. objective) and task type (controlled vs. spontaneous) by looking at where effect sizes went beyond zero for between-group contrasts or the upper-range of the comparison groups' 95% *CIs* ($d > 0.38$) for within-group contrasts.

Table 7

Summary of Effect Sizes (Cohen's d) according to Measurement Focus, Scoring Methods and Task Types

Proficiency dimension	Global		Specific			
Scoring method	Human intuition		Expert rating		Acoustic analyses	
Task/knowledge type	Controlled production	Spontaneous production	Controlled production	Spontaneous production	Controlled production	Spontaneous production
A. Between-group						
No. of primary studies	6	4	19	3	17	6
Ratio (out of 39 studies)	15.3%	10.2%	48.7%	7.6%	43.5%	15.3%
No. of contrasts	8	4	28	4	20	4
Ratio (out of 68 contrasts)	11.7%	5.8%	41.1%	5.8%	29.4%	5.8%
M	0.33	0.73	0.75	0.40	0.84	0.24
SE	0.22	0.24	0.15	0.26	0.18	0.14
Lower - 95% *CIs*	-0.18	-0.03	0.42	-0.44	0.45	-0.22
Upper – 95% *CIs*	0.85	1.50	1.07	1.26	1.24	0.71
B. Within-group						
No. of primary studies	10	15	27	9	25	10
Ratio (out of 67 studies)	14.9%	22.3%	40.2%	13.4%	37.3%	14.9%
No. of contrasts	11	16	41	10	29	12
Ratio (out of 119 contrasts)	9.2%	13.4%	34.4%	8.4%	24.3%	10.8%
M	0.85	0.51	0.69	0.84	0.76	0.62
SE	0.24	0.09	0.12	0.36	0.13	0.21
Lower - 95% *CIs*	0.30	0.30	0.45	0.05	0.49	0.13
Upper – 95% *CIs*	1.40	0.71	0.94	1.64	1.02	1.10

Focus of Measurement. Although our general analyses of N = 77 pronunciation teaching studies noted the significant effects of instruction in both within- and between-group contrasts, our examination of effect sizes at a finer level (global vs. specific measures) showed slightly different observations. In the moderator analyses presented below, we determined the presence and absence of statistically significant differences in the following manner. The between-group comparisons concerned whether 95% CIs passed through zero; and the within-group comparisons related to whether 95% CIs went beyond the upper-range of the comparison groups' 95% *CIs* (d > 0.38).[6]

With respect to between-group contrasts, results showed the relatively small-to-medium effects of pronunciation teaching on the development of learners' controlled production of specific L2 pronunciation (segmental/suprasegmental accuracy) [*CI* = 0.42, 1.24]. Turning to global L2 pronunciation (e.g., comprehensibility), however, the differences between experimental and comparison groups in pronunciation teaching research fails to reach statistical significance. According to the 95% CIs, pronunciation teaching gains could be below zero (d < 0), when evaluated from perspectives of global proficiency in both task contexts. This indicates that a degree of ambiguity remains regarding the relationship between pronunciation teaching and global L2 pronunciation proficiency.

With respect to within-group contrasts, the *95% CIs* of d values ranged from small to medium, when primary studies expounded its effectiveness on specific L2 pronunciation proficiency at a controlled speech level [*CI* = 0.45, 1.02]. In contrast, the lower end of *95% CIs* overlapped with the *95% CIs* of the control group (0.24-0.38), when learner gains were analyzed through global measures or spontaneous measures. The results suggest that the effectiveness of pronunciation teaching could be limited to L2 learners' acquisition of specific segmental and suprasegmetal features at a controlled speech level.

Scoring Method. In between- and within-group contrasts alike, the amount of pronunciation teaching effectiveness on segmental and suprasegmental accuracy development does not appear to differ significantly across both controlled and spontaneous production contexts according to the scoring method (expert ratings or acoustic analyses). In the case of expert rating of specific L2 pronunciation proficiency, however, we note the relatively large standard errors especially when it comes to speech elicited via spontaneous tasks (*SEs* = .25 and .36 for between- and within-group, respectively). Comparatively, the standard errors of acoustic analyses appeared to be consistent regardless task conditions (*SEs* = .18-.24 for between-group; .13-.21 for within-group). Together with our analysis of reliability estimates reported earlier, it is reasonable to say that the expert rating approach may result in relatively stable and more varied evaluations, when it is used to assess L2 segmental and suprasegmental accuracy (rather than global impressions) of spontaneous (rather than controlled) L2 speech.

[6] In this paper, we did not provide any interpretations about the degree of overlap between 95% *CIs* of effect sizes, because we do not have any clear guidelines regarding how to do so specific to instructed SLA research. Rather, we restricted our analyses and discussion to the presence and absence of significant differences without any mention of the degree of differences, meaningfulness, and reliability.

Task Type. In between-group designs, the effect of task modality is observed more clearly. Spontaneous measures produce significantly smaller effects than measures of controlled production. The 95% CIs for all three subgroups involving spontaneous speech cross zero, indicating that the differences between experimental and comparison groups in pronunciation teaching research could be statistically unstable when measured in this fashion. For within-group contrasts, the 95% CIs demonstrated that the amount of improvement following pronunciation teaching [CI = 0.05, 1.64] demonstrated substantial overlap with the control group [CI = 0.24, 0.38], when learners' pronunciation on a specific feature was assessed in the context of spontaneous production tasks.

Discussion

In this study, pronunciation instruction is defined as provision of explicit metalinguistic information as to articulatory (how to produce) and/or auditory (how to perceive) aspects of new segmental and suprasegmental features in an L2. The current study set out to propose a model as to how to measure such instructed L2 pronunciation learning in classroom settings and for how to interpret relevant findings—i.e., Framework for L2 Pronunciation Measurement. This model, the first that we are aware of in the realm of L2 pronunciation, distinguishes between different (a) foci of measurement, (b) scoring methods, and (c) task types, producing six unique classifications at the finest level of granularity. In the context of N = 77 primary studies with a pre- and posttest design, we first conducted a research synthesis to survey how extant pronunciation literature has measured the effectiveness of instruction on six different dimensions of L2 pronunciation proficiency and knowledge as specified in our model. To shed light on the methodological innovation and reform especially after the publication of Levis's (2005) special issue in *TESOL Quarterly*, sub-group analyses were also conducted on publications up until 2007 and from 2008 (k = 22 between 1982 and 2007, 55 between 2008 and 2017, respectively). As a means to provide empirical support for this classification scheme, and demonstrate the state-of-the-art status of pronunciation teaching research through the proposed model, the current study then meta-analyzed the effects of instruction. More specifically, we aimed to re-examine the pedagogical potential of pronunciation teaching according to different constructs (global vs. specific) and scoring methods (expert rating vs. acoustic analyses) of L2 pronunciation proficiency and knowledge (controlled vs. spontaneous). We also applied this same framework using reliability generalization meta-analysis as a means to examine estimates of measurement error as a function of the target constructs and measures in our model.

Overall, three crucial findings emerged. First, whereas primary studies have increasingly looked at the multifaceted nature of instructed L2 pronunciation learning by incorporating multiple analyses (human judgements, acoustic analyses) and tasks (controlled, spontaneous) over the past 10 years, very few scholars have probed the impact of instruction on both global and specific L2 proficiency. Second, evidence that participants significantly and substantially benefited from pronunciation instruction is especially robust when outcome measures focused on their more controlled and monitored specific L2 segmental and suprasegmental proficiency. In contrast, the results indicated a lack of significant effects of instruction when outcome measurements focused on global rather than specific proficiency, and when its gains were measured via spontaneous tasks. Third, comparing expert rating (subjective analyses) vs. acoustic analyses (objective analyses), the results of reliability meta-analysis and standard error indicated that the effectiveness of instruction could be slightly more varied (alpha = .76; *SE* =.35, .26) when the spontaneous speech samples were evaluated through experts' impressionistic judgements as opposed to acoustic analyses (*SE* =.21, .14).

Taken together, we view the results of the present study as providing empirical support for the usefulness of the Framework for L2 Pronunciation Measurement as a tool to synthesize a great deal of variability in study outcomes and present a more lucid, meaningful and unified understanding of instructed L2 pronunciation learning. We explain this interpretation and interpret many of its substantive and methodological implications in detail below.

Focus of Measurement

The first broad conclusion we would like to make is that the efficacy of pronunciation teaching remains relatively unclear when specific features are the focus of instruction (e.g., segmentals, suprasegmentals) and gains are measured globally (e.g., comprehensibility, accentedness, intelligibility, fluency). According to the results of our research synthesis, while most of the primary studies examined the effect of instruction on specific L2 pronunciation proficiency improvement (70.1%), some pronunciation teaching researchers *have* been interested in furthering the extent to which pronunciation teaching can ultimately lead to positive influences on the global constructs of L2 pronunciation proficiency (e.g., comprehensibility) (10.3%).

Though relatively infrequent, some pronunciation teaching studies (19.4%) approached the effects of pronunciation teaching only through global L2 pronunciation measures (e.g., comprehensibility). The number of the latter type of research has increased over the past 10 years (13.6% → 23.6%). This is surprising, given that these studies appeared to concern only the ultimate goal of pronunciation teaching (the impact of instruction on global L2 pronunciation proficiency development); but not the actual outcomes of the intervention that they had delivered (the extent to which explicit instruction could be facilitative of L2 segmental and suprasegmental learning).

As shown by the effect size analyses, pronunciation teaching (i.e., teaching segmental and suprasegmental accuracy) was indeed found to be facilitative of, in particular, "specific" L2 pronunciation proficiency at a controlled speech level; however, its impact on "global" L2 pronunciation remains unclear under both controlled and spontaneous task conditions (in that their lower CIs went below 0.38). The results suggest that pronunciation teaching may be a helpful, but not generally sufficient, condition for global L2 pronunciation learning. That is, it is true that accurate segmental and suprasegmental accuracy comprises a primary linguistic cue that listeners rely on during any global judgements (e.g., accentedness, comprehensibility) (Crowther et al., 2015; Kang et al., 2010; Saito et al., 2017). However, even if we provide explicit instruction on certain segmental and suprasegmental features, it does not necessarily improve global L2 pronunciation proficiency.

This is arguably because listeners also pay attention to a number of factors beyond phonological information while assessing global proficiency. During L2 comprehensibility judgements, for example, listeners tend to attend to the phonological, temporal, lexical and grammatical dimensions of language in order to collect as much linguistic information as possible from accented speech (Crowther et al., 2015; Isaacs et al., 2017; Saito et al., 2017); this tendency reflects the main objective of L2 comprehensibility assessment, which is to obtain a general picture of what talkers intend to convey as quickly as possible (Munro & Derwing, 1995). For another example, when asked to rate for global foreign accentedness, it has been shown that listeners take into account not only phonological accuracy (Riney et al., 2000) but also lexicogrammatical sophistication (Ruivivar & Collins, 2018).

In conjunction with the importance of matching the foci of instruction with appropriate outcome measures, as stressed by Norris and Ortega (2012), it is advisable for future pronunciation teaching researchers to explicitly clarify what the instruction actually aims to teach (segmental and suprasegmental accuracy), and how learner improvement is measured (via adequate segmental and suprasegmental measures). This is especially true if such studies highlight the influence of pronunciation teaching on both the specific *and* global constructs of L2 pronunciation proficiency. Researchers conducting this kind of research are thus strongly recommended to carefully explain the logical connections between (a) teaching specific pronunciation features; (b) its immediate effect on segmental/suprasegmental development; and (c) the ultimate impact of pronunciation teaching on improved global pronunciation proficiency. In particular, researchers should address the indirect link between pronunciation teaching and global L2 pronunciation proficiency because of the potential influence of other non-phonological factors (e.g., lexicogrammar errors). It is probable that improving global L2 pronunciation proficiency may require comprehensive instructional treatment targeting the acquisition of a wide range of pronunciation, fluency, vocabulary and grammar features in a complementary fashion (see Isaacs et al., 2017 for a list of linguistic correlates of L2 comprehensibility).

Scoring Method

The primary studies in the current dataset exhibit a preference for impressionistic judgements over acoustic analyses. This is arguably because what pronunciation teaching researchers seek to expound is whether pronunciation teaching can lead to “perceptible” change in learner speech at a macro level, rather than whether pronunciation teaching can change certain acoustic properties of speech at a micro level (Thomson & Derwing, 2015). According to the results of effect size analyses, we did not find any significantly differential effects of the scoring methods in any contexts (subjective vs. objective). Given that our reliability estimates and standard error analyses hinted at some potential patterns, however, we would like to discuss them but as *tentative* observations.

First, intuitive assessments of instructional gains in global pronunciation demonstrated relatively high agreement (.87-.93), suggesting that even naïve listeners have a shared notion of perceived comprehensibility, intelligibility, perceived fluency and accentedness (Derwing & Munro, 2005; see similarly high estimates in the broader domain of

"L2 pronunciation" sampled by Plonsky & Derrick, 2016)[7]. Such impressionistic judgements resulted in less stable evaluations, however, when raters assessed for specific segmental and suprasegmental accuracy, especially at a spontaneous speech level (.76). Furthermore, the results of standard error analyses suggested that such expert raters' evaluation of spontaneous L2 segmentals and suprasegmentals may also be subject to variation (*SEs* = .26 for between-group comparisons, .36 for within-group comparisons) relative to acoustic analyses (*SEs* = .14 for between-group comparisons, .21 for within-group comparisons).

The results here indicate that raters may have difficulty in assessing spontaneous L2 speech in a consistent manner. While making impressionistic, global judgments of spontaneous speech, raters are found to take into account both phonological and non-phonological factors (e.g., lexical, pragmatic and contextual information) (Broersma & Cutler, 2008); their evaluations are likely influenced by their previous listening experience with particular foreign accents (e.g., familiarity) (Bent & Bradlow, 2008); and they use different acoustic information to analyze the same sound according to different rating constructs (accuracy vs. intelligibility) and talkers' proficiency (beginner vs. intermediate vs. advanced) (e.g., Saito, 2013; Saito & Munro, 2014).

Once again, our discussion here should be considered as tentative at best, since we failed to find any statistically significant findings due to different types of scoring method. However, the results of reliability estimate and standard error analyses at least pointed out a possibility that using the different scoring methods (subjective, objective) may require different types of operationalization and interpretations, especially when researchers are interested in measuring instructional gains through spontaneous (rather than controlled) speech tasks. It is promising that the results of our synthesis showed that more researchers have been adopting both subjective and objective analyses in more recent publications (from 2008) (10.5% → 20.9%). Although the choice of scoring methods may not directly impact the amount of instructional gains observed, using both the subjective/objective analyses can tap into two complementary aspects of L2 pronunciation proficiency, providing a more comprehensive picture of instructed L2 pronunciation development at perceptible and fine-grained levels (see Saito & van Poeteren, 2018 for more relevant discussion on the relationship between subjective and objective analyses across different task conditions).

Task Type

Our final conclusion is that pronunciation teaching is most effective when it targets specific pronunciation features (segmentals, syllables, prosody, utterance fluency) and when instructional gains are measured via controlled tasks (word and sentence reading). Our findings suggest that pronunciation teaching can directly promote the development of L2 learners' explicit, controlled and specific pronunciation proficiency. Yet, the robustness and transferability of such instructional gains may remain unclear especially under more

[7] While reliability in this area is relatively high, there has been some evidence that despite high intraclass correlation coefficients, listeners are not necessarily rating in the same way (see Isbell, 2018). Although little attention has been given to the process of global L2 pronunciation judgements, this topic should be further examined in the future.

spontaneous contexts. The number of studies using the more spontaneous tests was low once our study sample was broken down into different types of scoring, different proficiency dimensions, and in the studies with between-group comparisons. However, significant task effects were evidenced as fairly consistent phenomena in the results of the between-group and within-group comparisons.

For the between-group comparisons, all three spontaneous production columns have *CIs* that pass through zero (see Table 7). For the within-group comparisons, this interpretation is evidenced by our finding that effect sizes on spontaneous tests fell below the *CIs* of the control groups' pre-post effects, whereas those from controlled tests almost always fell above the control groups' *CIs*. The trends that our analysis has revealed (controlled > spontaneous) are in line with the extensive literature on instructed SLA indicating that the positive effects of explicit instruction are tied to the initial stages of SLA (noticing, pattern identification, restructuring, error avoidance). However, we also suggest that learners be given opportunities to proceduralize and automatize the target features (DeKeyser, 2017; Gatbonton & Segalowitz, 2005; Spada & Tomita, 2010).The relatively strong effects of pronunciation teaching on controlled (but not on spontaneous/automatized) knowledge could also be ascribed to the methodological problems that we identified in our synthesis of primary pronunciation teaching studies—i.e., scholars' over-reliance on controlled tasks. Whereas most of the pronunciation teaching studies relied exclusively on controlled tasks (57.1%), very few studies actually adopted both controlled and spontaneous tasks (18.1%). This methodological bias prevents us from making any rock-solid conclusions with regard to the effect of pronunciation teaching on different types of L2 pronunciation knowledge (controlled vs. spontaneous/automatized).

The same phenomenon (lack of spontaneous measures) was observed when Norris and Ortega (2000) meta-analyzed the effects of L2 grammar instruction. This methodological practice was also severely criticized in Doughty's (2003) narrative review and, later, taken up again in Mackey and Goo's (2007) meta-analysis of the effects of L2 interaction as well as in Goo, Granena, Yilmaz, and Novella's (2015) meta-analytic replication of Norris and Ortega (2000). Although many L2 grammar teaching studies supported the effectiveness of form-focused (rather than meaning-oriented) instruction, Doughty cautioned that most of the outcome measures in these studies may have simply reflected how much L2 learners had accumulated metalinguistic awareness and knowledge. Indeed, few of these studies examined the extent to which instruction could help L2 learners actually access the target language when using it spontaneously for communicative purposes.

Such critiques eventually resulted in methodological renovations in the field, with a growing number of L2 scholars making efforts to scrutinize the impact of instruction on various modes of SLA by adopting both controlled and spontaneous tasks (e.g., 16% of primary studies in Norris & Ortega [2000] vs. 50% in Spada and Tomita [2010]). According to the results of our research synthesis, similar methodological innovation has taken place in pronunciation teaching research. As certain scholars have recommended (e.g., Saito, 2012), more primary studies have tended in recent years to include both controlled and spontaneous speech tasks (13.6% → 29.0%).

Here, we would like to reiterate that controlled and spontaneous L2 pronunciation performance are essentially two different phenomena, and thus need to be assessed and interpreted separately, as exhibited in our proposed model. Previous investigations in classroom (e.g., Rau et al., 2009) and naturalistic (e.g., Saito & Brajot, 2013) settings have revealed that once learners start making efforts to acquire new sounds, their performance is largely susceptible to variations in task conditions, with performance being more targetlike at a controlled than at a spontaneous speech level.

Both researchers and practitioners would agree that pronunciation teaching can greatly help L2 learners attain enhanced awareness, noticing and understanding of new sounds via explicit phonetic instruction (Derwing & Munro, 2005). In the long run, pronunciation teaching should also induce L2 learners to engage in certain communicative activities using the newly acquired knowledge (e.g., speaking language for meaning with correct pronunciation forms) (i.e., proceduralization), and ultimately attain highly accurate, fluent and effortless use of L2 pronunciation (i.e., automatization) (DeKeyser, 2017). To this end, we strongly emphasize that future studies should adopt both controlled and spontaneous tasks to capture the multifaceted nature of gains resulting from pronunciation teaching. This methodological decision will allow us to examine, in particular, the impact of instruction on L2 learners' pronunciation abilities during their processing of language for meaning in future communicative settings.

However, we need to acknowledge that our proposed framework does not tap into a more "fine-grained" distinction between different types of knowledge and corresponding tasks, nor did the studies within our study sample. For example, spontaneous knowledge, featured in our proposed model, covers a wide range of behavioral phenomena, whereby L2 learners can quickly access what they have learned from instruction, but with varied degrees of awareness (Spada & Tomita, 2010). This more spontaneous knowledge could be considered equivalent to explicit automatized knowledge, referred to as "a body of conscious linguistic knowledge including different levels of automatization" (e.g., Suzuki, 2017, p. 1230). Yet, such spontaneous and automatized knowledge has been claimed to be different from implicit knowledge which is acquired when L2 learners are not aware of what they are learning nor what they have acquired (DeKeyser, 2017). Focusing on naturalistic, advanced-level L2 learners (rather than beginner/intermediate L2 learners in classroom settings), the existing research on automatized vs. implicit knowledge has provided much information especially to theory building on the underlying mechanisms of human language acquisition (i.e., roles of explicit vs. implicit cognition); in these studies, it has remained controversial what type of task modality can be sensitive enough to capture the theoretically vital distinction between automatized and implicit knowledge (e.g., see Plonsky et al., in press; Suzuki & DeKeyser, 2017; Vafaee, Suzuki, & Kachisnke, 2017).

In the proposed framework, this implicit knowledge is not included, since its discussion (i.e., explicit or implicit) is not directly relevant to pronunciation teaching studies, where both practitioners and researchers are more interested in the role of instruction in L2 pronunciation development *with* learner awareness (using language at both controlled and spontaneous levels). Echoing DeKeyser's (2017) skill acquisition framework for instructed SLA, our model highlights the *declarative-procedural-automatized* (instead of *explicit-implicit*) distinction that better reflects what teachers and students do—i.e., provision of declarative knowledge followed by proceduralization and automatization activities. Different from the explicit-implicit distinction, where scholars have debated the validity of outcome measures, there is a methodological consensus in cognitive psychology and SLA alike that declarative, procedural and automatized knowledge can be measured via single and dual task conditions, respectively. At the same time, we also call for more future research which will further theorize, disentangle, test and elaborate on (a) what comprises explicit and implicit phonetic knowledge especially beyond classroom L2 pronunciation learning; and (b) what kinds of outcome measures can tap into learners' proceduralization and automatization of L2 pronunciation with varying degrees of awareness. Such future research will eventually help us understand the complex mechanisms underlying advanced L2 learners' long-term pronunciation attainment, nativelikeness and individual differences especially in naturalistic settings (Suzuki & DeKeyser, 2017).

Conclusion

On the whole, the results of our focused meta-analysis on 77 primary studies lend empirical support for use of the proposed model (i.e., Framework for L2 Pronunciation Measurement) to synthesize methodological practices and empirical evidence in pronunciation teaching research. These results provide tentative suggestions that pronunciation teaching may differentially influence the development of L2 pronunciation proficiency and knowledge depending on how it is evaluated from three different angles: (a) global vs. specific constructs; (b) subjective vs. objective analyses; and (c) controlled vs. spontaneous knowledge. On the one hand, pronunciation teaching can be beneficial at a controlled level, as providing explicit phonetic information enables learners to notice and practice the accurate production of L2 segmental, syllabic, prosodic and temporal features in a careful fashion. On the other hand, our results cast doubt on (a) whether and to what degree pronunciation teaching can subsequently lead to perceptible changes in learners' relatively spontaneous and automatized pronunciation performance; and (b) whether pronunciation teaching can ultimately impact global pronunciation proficiency (e.g., comprehensibility). In general, when it comes to the analysis of specific segmental and suprasegmental L2 pronunciation accuracy, subjective measures (expert judgements) seem to involve more variability than objective measures (acoustic analyses), especially when they are applied to spontaneous speech tasks. This is arguably because the former approach involves raters who inevitably take into account not only phonological, but also other factors which might affect the perception and judgment of speech quality (e.g., contextual factors, listener backgrounds). Such non-construct relevant variance (i.e., noise), when present in the data, also obscures the signal being sought and attenuates (reduces) observed effects. However, it is important to remind readers that the discussion here is based on the relatively small sample size (k = 3 and 9 primary studies for between- and within-group comparisons).

We would like to argue that it is theoretically, methodologically and pedagogically crucial for future pronunciation teaching researchers to make it clear which constructs of L2 pronunciation proficiency their pronunciation teaching treatments purport to target in the short (e.g., specific segmental and suprasegmental features) and long run (e.g., comprehensibility); and how the effectiveness of pronunciation teaching is tested (controlled vs. spontaneous levels) and analyzed (subjectively vs. objectively). In accordance with our proposed framework, we recommend that pronunciation teaching researchers make predictions, interpret results, and connect their discussion with overall theoretical accounts of instructed SLA. We also encourage such interpretations to take into account previous findings based on comparable investigations (for similar arguments, see Loewen, 2014). Both the effect size and reliability estimates in the present study provide one potentially useful source for doing so.

To close, and with a view towards providing further guidance for future research, we propose the following recommendations for overcoming the limitations raised in this current study. First, it has been argued that the ultimate goal of teaching specific pronunciation proficiency is to enhance L2 learners' overall comprehensibility (one form of global pronunciation proficiency) (e.g., Derwing & Munro, 2005). Although the current study did not evidence particularly strong relations between pronunciation teaching and global pronunciation proficiency, future studies should further pursue efforts to enhance the benefits of pronunciation teaching by carefully elaborating the content of the pronunciation teaching treatments involved. For example, we recommend that scholars select pronunciation features which greatly affect listeners' comprehension (Derwing et al., 1998), and integrate pronunciation teaching into vocabulary, grammar and pragmatics lessons so that L2 learners can learn various aspects of language which are equally relevant to listeners' overall judgments of L2 comprehensibility (Saito et al., 2016).

Second, given we noted more variance in expert rating compared with acoustic analyses of spontaneous L2 speech, researchers need to interpret any emerging findings with caution and flexibility. For example, the lack of statistical significance does not necessarily reject the effects of pronunciation teaching (rather indicating individual differences in human perception of spontaneous speech). In this regard, we must wait for future studies to further examine the complex relationship between such subjective and objective scoring methods—i.e., how raters use acoustic information to perceive specific segmental and suprasegmental features in various task, phonetic and interlocutor contexts (cf. Saito & van Poeteren, 2018 for segmentals; Bosker et al., 2013 for fluency).

Third, including both controlled and spontaneous tasks is crucial to obtaining a better and more detailed picture of the effectiveness of pronunciation teaching. As recent studies have expounded a list of communicatively oriented practice activities for L2 pronunciation development (see Mora & Levkina, 2017), the impact of different types of pronunciation teaching on controlled and automatized knowledge should be adequately assessed via controlled and spontaneous tasks. We consider "type-of-instruction" as one promising direction for future pronunciation teaching research. Theoretically and pedagogically intriguing options include form vs. meaning-oriented instruction (e.g., Saito, 2012), articulatory vs. auditory-based training (Sakai, 2016), audio-only vs. audiovisual (Hardison, 2003), and input- and output-based feedback (Gooch, Saito, & Lyster, 2016).

Last, it is clear from the reliability generalization meta-analysis that greater consideration is needed with respect to the psychometric properties of pronunciation teaching measures, as in much of applied linguistics. At the very least and as a matter of course, estimates of reliability should be reported and discussed to allow for informed interpretations of primary research. This is not a new recommendation. Despite its long-standing inclusion in the guidelines of most learned societies and journals (e.g., *Language Learning*; see Norris, Plonsky, Ross, & Schoonen, 2015), numerous syntheses have noted the lack of rigor and transparency in this area at levels comparable to the domain of the present study (Derrick, 2016; Plonsky & Gass, 2011). We would also encourage researchers in pronunciation teaching to consider going one step further by applying psychometric 'corrections' to their analyses when suboptimal reliability estimates are obtained (see Muchinsky, 1996; for recent examples in applied linguistics, see Llosa & Malone, in press, and Teimouri, 2017). These procedures are computationally straightforward and potentially very useful in that they allow the researcher to model the relationships of interest in the absence of attenuation (reduction) of effects due to measurement error.

References

Barriuso, T. A., & Hayes-Harb, R. (2018). High Variability Phonetic Training as a Bridge from Research to Practice. *CATESOL Journal*, *30*, 177-194.

Bachman, L. F. (1988). Language testing-SLA interfaces. *Annual Review of Applied Linguistics, 9*, 193-209. DOI: https://doi.org/10.1017/S0267190500000891

Bachman, L. F., & Cohen, A. D. (1998). *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press.

Bent, T., & Bradlow, A. (2003). The interlanguage intelligibility benefit. *Journal of Acoustical Society of America, 114*, 1600–1610. DOI: https://doi.org/10.1121/1.1603234

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

Bohn, O.-S., & Flege, J. (1997). Perception and production of a new vowel category by second-language learners. In A. James & J. Leather (Eds.), *Second-language speech: Structure and process* (pp. 53–74). Berlin: Mouton de Gruyter.

Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T., & De Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing, 30*, 159–175. DOI: https://doi.org/10.1177/0265532212455394

Broersma, M., & Cutler, A. (2008). Phantom word activation in L2. *System*, *36*, 22-34. DOI: https://doi.org/10.1016/j.system.2007.11.003

Celce-Murcia, M., Brinton, D., Goodwin, J., & Griner, B. (2010). *Teaching pronunciation: A course book and reference guide*. Cambridge: Cambridge University Press.

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

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2015). Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL Quarterly, 49*, 814-837. DOI: https://doi.org/10.1002/tesq.203

DeKeyser, R. M. (2017). Knowledge and skill in ISLA. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of second language acquisition* (pp. 15-32). New York and London: Routledge.

Derrick, D. J. (2016). Instrument reporting practices in second language research. *TESOL Quarterly, 50,* 132-153. DOI: https://doi.org/10.1002/tesq.217

Derwing, T., & Munro, M. (2005). Second language accent and pronunciation teaching: A research–based approach. *TESOL Quarterly*, *39*, 379–397. DOI: https://doi.org/10.2307/3588486

Derwing, T. M., Munro, M. J., Foote, J. A., Waugh, E., & Fleming, J. (2014). Opening the window on comprehensible pronunciation after 19 years: A workplace training study. *Language Learning*, *64*, 526–548. DOI: https://doi.org/10.1111/lang.12053

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

Derwing, T.M., Rossiter, M.J., Munro, M.J. & Thomson, R.I. (2004). L2 fluency: Judgments on different tasks. *Language Learning, 54,* 655–679. DOI: https://doi.org/10.1111/j.1467-9922.2004.00282.x

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.

Flege, J. (2016, June). *The role of phonetic category formation in second language speech acquisition*. Plenary address delivered at New Sounds, Aarhus, Denmark.

Flege, J., Munro, M, & MacKay, I. R. A. (1995). Factors affecting degree of perceived foreign accent in a second language. *Journal of the Acoustical Society of America, 97*, 3125–3134. DOI: https://doi.org/10.1121/1.413041

Flege, J., Takagi, N., & Mann, V. (1995). Japanese adults learn to produce English /ɹ/ and /l/ accurately. *Language and Speech, 38*, 25–55. DOI: https://doi.org/10.1177/002383099503800102

Gatbonton, E. & Segalowitz, N. (2005). Rethinking the communicative approach: A focus on accuracy and fluency. *Canadian Modern Language Review, 61*, 325–353. DOI: https://doi.org/10.3138/cmlr.61.3.325

Gick, B., Bernhardt, B., Bacsfalvi, P., & Wilson, I. (2008). Ultrasound imaging applications in second language acquisition. *Phonology and Second Language Acquisition*, *36*, 315-328.

Ginther, A., & Yan, X. (2018). Interpreting the relationships between TOEFL iBT scores and GPA: Language proficiency, policy, and profiles. *Language Testing*, *35*, 271-295. DOI: https://doi.org/10.1177/0265532217704010

Goo, J., Granena, G., Yilmaz, Y., & Novella, M. (2015). Implicit and explicit instruction in L2 learning: Norris & Ortega (2000) revisited and updated. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (pp. 443-482). Amsterdam: John Benjamins.

Gooch, R., Saito, K., & Lyster, R. (2016). Effects of recasts and prompts on L2 pronunciation development: Teaching English /r/ to Korean adult EFL learners. *System, 60,* 117-127. DOI: https://doi.org/10.1016/j.system.2016.06.007

Gorsuch, G. J. (2011). Improving Speaking fluency for international teaching assistants by increasing input. *TESL-EJ*, *14*(4), n4.

Harding, L. (2017). Validity in pronunciation assessment. In O. Kang, & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 30-48). London: Routledge.

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

Isaacs, T., Trofimovich, P., & Foote, J. A. (2017). Developing a user-oriented second language comprehensibility scale for English-medium universities. *Language Testing.* doi: 10.1177/0265532217703433

Isaacs, T., Trofimovich, P., Yu, G., & Muñoz Chereau, B. (2015). Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS pronunciation scale. *IELTS research reports online series*, *4*.

Isbell, D. (2018). Assessing pronunciation for research purposes with listener-based numerical scales. In O. Kang & A. Ginther (Eds.), *Assessment of second language pronunciation* (pp. 89-112). New York: Routledge.

Kang, O., Rubin, D., Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of English language learner proficiency in oral English, *Modern Language Journal, 94,* 554–566. DOI: https://doi.org/10.1111/j.1540-4781.2010.01091.x

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. DOI: https://doi.org/10.3138/cmlr.64.3.459

Kennedy, S., Blanchet, J., & Trofimovich, P. (2014). Learner pronunciation, awareness and instruction in French as a second language. *Foreign Language Annals*, *47*, 79-96. DOI: https://doi.org/10.1111/flan.12066

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience, 5,* 831–843.

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. DOI: https://doi.org/10.1017/S0142716405050150

Lambert, C., Kormos, J., & Minn, D. (2017). Task repetition and second language speech processing. *Studies in Second Language Acquisition*, *39*, 167–196. DOI: https://doi.org/10.1017/S0272263116000085

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning, 65*, Supp. 1, 127-159. DOI: https://doi.org/10.1111/lang.12115

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

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

Levis, J. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, *39*, 367–377. DOI: https://doi.org/10.2307/3588485

Levis, J. M., Sonsaat, S., Link, S., & Barriuso, T. A. (2016). Native and nonnative teachers of L2 pronunciation: Effects on learner performance. *TESOL Quarterly*, *50*, 894-931. DOI: https://doi.org/10.1002/tesq.272

Lin, Y. H. (2003). Interphonology variability: Sociolinguistic factors affecting L2 simplification strategies. *Applied Linguistics*, *24*, 439-464. DOI: https://doi.org/10.1093/applin/24.4.439

Llosa, L., & Malone, M. E. (in press). Comparability of students' writing performance on TOEFL iBT and in required university writing courses. *Language Testing*. DOI: https://doi.org/10.1177/0265532218763456

Loewen, S. (2014). *Introduction to instructed second language acquisition*. New York, NY:Routledge.

Lyster, R., & Saito, K. (2010). Corrective feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition, 32*, 265-302. DOI: https://doi.org/10.1017/S0272263109990520

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. 407-451). New York: Oxford University Press.

Major, R. (2001) *Foreign accent: The ontogeny and phylogeny of second language phonology*. Mahwah, NJ: Lawrence Erlbaum Associates.

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning*, *68*, 321-391. DOI: https://doi.org/10.1111/lang.12286

Marsden, E. J., Thompson, S., & Plonsky, L. (2018). A methodological synthesis of self-paced reading in second language research: Methodological synthesis of SPR tests. *Applied Psycholinguistics*, 861-904. DOI: https://doi.org/10.1017/S0142716418000036

Marsden, E. J., & Torgerson, C. J. (2012). Single group, pre- and post- research designs: Some methodological concerns. *Oxford Review of Education*, *38*, 583-616. DOI: https://doi.org/10.1080/03054985.2012.731208

Martinsen, R., Montgomery, C., & Willardson, V. (2017). The Effectiveness of Video-Based Shadowing and Tracking Pronunciation Exercises for Foreign Language Learners. *Foreign Language Annals*, *50*, 661-680. DOI: https://doi.org/10.1111/flan.12306

Mora, J. C., & Levkina, M. (2017). Task-based pronunciation teaching and research: Key issues and future directions. *Studies in Second Language Acquisition*, *39*, 381-399. DOI: https://doi.org/10.1017/S0272263117000183

Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement, 56*, 3–75. DOI: https://doi.org/10.1177/0013164496056001004

Munro, M., & Derwing, T. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech*, *38*, 289–306. DOI: https://doi.org/10.1177/002383099503800305

Munro, M., & Derwing, T. (2011). The foundations of accent and intelligibility in pronunciation research. *Language Teaching, 44*, 316–327. DOI: https://doi.org/10.1017/S0261444811000103

Nagle, C. (2018). Perception, production, and perception-production: Research findings and implications for language pedagogy. *World Languages and Cultures Publications*. 171.

Norris, J., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning, 50,* 417–528. DOI: https://doi.org/10.1111/0023-8333.00136

Norris, J., & Ortega, L. (2012). Assessing learner knowledge. In S. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 716-761). New York: Routledge.

Norris, J. M., Plonsky, L., Ross, S. J., & Schoonen, R. (2015). Guidelines for reporting quantitative methods and results in primary research. *Language Learning, 65*, 470-476. DOI: https://doi.org/10.1111/lang.12104

Offerman, H. M., & Olson, D. J. (2016). Visual feedback and second language segmental production: The generalizability of pronunciation gains. *System*, *59*, 45-60. DOI: https://doi.org/10.1016/j.system.2016.03.003

Oh, G. E., Guion-Anderson, S., Aoyama, K., Flege, J. E., Akahane-Yamada, R., & Yamada, T. (2011). A one-year longitudinal study of English and Japanese vowel production by Japanese adults and children in an English-speaking setting. *Journal of phonetics*, *39*, 156-167. DOI: https://doi.org/10.1016/j.wocn.2011.01.002

Paradis, M. (2009). *Declarative and procedural determinants of second languages*. Amsterdam: Benjamins.

Parlak, Ö., & Ziegler, N. (2017). The impact of recasts on the development of primary stress in a synchronous computer-mediated environment. *Studies in Second Language Acquisition*, *39*, 257-285. DOI: https://doi.org/10.1017/S0272263116000310

Piske, T., Flege, J., MacKay, & Meador, D. (2011). Investigating native and non-native vowels produced in conversational speech. In M. Wrembel, M. Kul & K. Dziubalska-Kołaczyk (Eds.), *Achievements and perspectives in the acquisition of second language speech: New Sounds 2010* (pp. 195–205). Frankfurt am Main: Peter Lang.

Plonsky, L. (2017). Quantitative research methods. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 505-521). New York, NY: Routledge.

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *Modern Language Journal*, *100*, 538-553. DOI: |https://doi.org/10.1111/modl.12335

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, *61*, 325-366. DOI: https://doi.org/10.1111/j.1467-9922.2011.00640.x

Plonsky, L., Marsden, E., Crowther, D., Gass, S., & Spinner, P. (in press). A methodological synthesis of judgment tasks in second language research. *Second Language Research.*

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

Rau, D., Chang, A., & Tarone, E. (2009). Think or sink: Chinese learners' acquisition of the voiceless interdental fricative. *Language Learning*, *59*, 581–621. DOI: https://doi.org/10.1111/j.1467-9922.2009.00518.x

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning*, *63*, 595–626. DOI: tps://doi.org/10.1111/lang.12010

Riney, T., Takada, M., & Ota, M. (2000). Segmentals and global foreign accent: The Japanese flap in EFL. *TESOL Quarterly, 34*, 711–737. DOI: https://doi.org/10.2307/3587782

Ruivivar, J., & Collins, L. (2018). The effects of foreign accent on perceptions of nonstandard Grammar: A pilot study. *TESOL Quarterly*, *52*, 187-198. DOI: 10.1002/tesq.374

Saito, K. (2012). Effects of instruction on L2 pronunciation development: A synthesis of 15 quasi-experimental intervention studies. *TESOL Quarterly*, 842-854. DOI: https://doi.org/10.1002/tesq.67

Saito, K. (2013). The acquisitional value of recasts in instructed second language speech learning: Teaching the perception and production of English /r/ to adult Japanese learners. *Language Learning, 63*, 499-529. DOI: https://doi.org/10.1111/lang.12015

Saito, K. (2018a). Advanced segmental and suprasegmental acquisition. In P. Malovrh & A. Benati (Eds.). *The handbook of advanced proficiency in second language acquisition* (pp. 282-303). Wiley Blackwell.

Saito, K. (2018b). Individual differences in second language speech learning in classroom settings: Roles of awareness in the longitudinal development of Japanese learners' English /r/ pronunciation. *Second Language Research*. DOI: https://doi.org/10.1177/0267658318768342

Saito, K., & Brajot, F. (2013). Scrutinizing the role of length of residence and age of acquisition in the interlanguage pronunciation development of English /r/ by late Japanese bilinguals. *Bilingualism: Language and Cognition, 16*, 847-863. DOI: https://doi.org/10.1017/S1366728912000703

Saito, K., & Lyster, R. (2012). Effects of form-focused instruction and corrective feedback on L2 pronunciation development of /r/ by Japanese learners of English. *Language Learning, 62*, 595-633. DOI: https://doi.org/10.1111/j.1467-9922.2011.00639.x|

Saito, K., & Munro, M. (2014). The early phase of /r/ production development in adult Japanese learners of English. *Language and Speech, 57*, 451-469. DOI: DOI: https://doi.org/10.1177/0023830913513206

Saito, K., Tran, M., Suzukida, Y., Sun, H., Magne, V., & Ilkan, M. (in press). How do second language listeners perceive the comprehensibility of foreign-accented speech? Roles of first language profiles, second language proficiency, age, experience, familiarity and metacognition. *Studies in Second Language Acquisition.*

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

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

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical profiles of comprehensible second language speech: The role of appropriateness, fluency, variation, sophistication, abstractness and sense relations. *Studies in Second Language Acquisition, 37*, 677-701. DOI: https://doi.org/10.1017/S0272263115000297

Saito, Y., & Saito, K. (2017). Differential effects of instruction on the development of second language comprehensibility, word stress, rhythm, and intonation: The case of inexperienced Japanese EFL learners. *Language Teaching Research, 21*, 589-608. DOI: https://doi.org/10.1177/1362168816643111

Sakai, M. (2016). *(Dis)Connecting perception and production: Training native speakers of Spanish on the English/i/-/ɪ/distinction* (Unpublished doctoral dissertation). Georgetown University, Washington, D.C.

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*, 187–224. DOI: https://doi.org/10.1017/S0142716417000418

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning, 60,* 263–308. DOI: https://doi.org/10.1111/j.1467-9922.2010.00562.x

Suzuki, Y. (2017). Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge. *Applied Psycholinguistics*, *38*, 1229-1261. DOI: https://doi.org/10.1017/S014271641700011X

Suzuki, Y., & DeKeyser, R. (2017). The interface of explicit and implicit knowledge in a second language: Insights from individual differences in cognitive aptitudes. *Language Learning*, *67*, 747-790. DOI: https://doi.org/10.1111/lang.12241

Teimouri, Y. (2017). L2 selves, emotions, and motivated behavior. *Studies in Second Language Acquisition, 3*9, 681-709. DOI: https://doi.org/10.1017/S0272263116000243

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

Trofimovich, P., & Baker, W. (2006). Learning second-language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition, 28,* 1–30. DOI: https://doi.org/10.1017/S0272263106060013

Trofimovich, P., Kennedy, S., & Blanchet, J. (2017). Development of Second Language French Oral Skills in an Instructed Setting: A Focus on Speech Ratings. *Canadian Journal of Applied Linguistics/Revue canadienne de linguistique appliquée*, *20*, 32-50. DOI: https://doi.org/10.7202/1042675ar

Vafaee, P., Suzuki, Y., & Kachisnke, I. (2017). Validating grammaticality judgment tests: Evidence from two new psycholinguistic measures. *Studies in Second Language Acquisition*, *39*, 59-95. DOI: https://doi.org/10.1017/S0272263115000455

Venkatagiri, H., & Levis, J. (2007). Phonological awareness and speech comprehensibility: An exploratory study. *Language Awareness, 16*, 263–277. DOI: https://doi.org/10.2167/la417.0

Wheeler, D. L., Vassar, M., Worley, J. A., & Barnes, L. L. B. (2011). A reliability generalization meta-analysis of coefficient alpha for the Maslach Burnout Inventory. *Educational and Psychological Measurement, 71*, 231–244. DOI: https://doi.org/10.1177/0013164410391579

Yan, X., & Ginther, A. (2017). Listeners and raters: Similarities and differences in evaluation of accented speech. In O. Kang, & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 67-88). London: Routledge.

Supporting Information-A: *N* = 77 Primary Studies Included in Meta-Analysis

AbuSeileek, A. F. (2007). Computer-assisted pronunciation instruction as an effective means for teaching stress. *The JALT CALL Journal, 3*, 3-14.

Akita, M. (2005). The effectiveness of a prosody-oriented approach in L2 perception and pronunciation training. *Academic studies, English Language and Literature, 53*, 1-22.

Alves, U., & Magro, V. (2011). Raising awareness of L2 phonology: Explicit instruction and the acquisition of aspirated /p/ by Brazilian Portuguese speakers. *Letras de Hoje, 46*, 71-80.

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

Bueno Alastuey, M. C. (2010). Synchronous-voice computer-mediated communication: Effects on pronunciation. *CALICO Journal, 28*, 1:1.

Champagne-Muzar, C. (1993). Second language accent: The role of the pedagogical environment, *IRAL, 31,* 143-160.

Chen, H., & Goswami, J. S. (2011). Structuring cooperative learning in teaching English pronunciation. *English Language Teaching, 4/3*, 26-32.

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

Crowther, D. (2016). Using what you know: Can cross-linguistic instruction improve L2 pronunciation. *Concordia Working Papers in Applied Linguistics, 6*. 27-45.

de Bot, K. (1983). Visual feedback of intonation I: Effectiveness and induced practice behavior. *Language and Speech*, *26*, 331-350.

de Bot, K., & Mailfert, K. (1982). The teaching of intonation: Fundamental research and classroom applications. *TESOL Quarterly*, *16*, 71-77.

de Jong, N., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, *61*, 533-568.

Derwing, T. M., Munro, M. J., Foote, J. A., Waugh, E., & Fleming, J. (2014). Opening the window on comprehensible pronunciation after 19 years: A workplace training study. *Language Learning*, *64*, 526-548.

Derwing, T. M., Munro, M. J., & Wiebe, G. (1997). Pronunciation Instruction for Fossilized Learners. Can it Help? *Applied Language Learning*, *8*, 217-35.

Elliott, A. R. (1995). Foreign language phonology: Field independence, attitude, and the success of formal instruction in Spanish pronunciation. *The Modern Language Journal*, *79*, 530-542.

Ferrier, L. J., Reid, L. N., & Chenausky, K. (1999). Computer-assisted accent modification: A report on practice effects. *Topics in Language Disorders*, *19*, 35-48.

Foote, J. A., & McDonough, K. (2017). Using shadowing with mobile technology to improve L2 pronunciation. *Journal of Second Language Pronunciation*, *3*, 34-56.

Galante, A., & Thomson, R. I. (2017). The effectiveness of drama as an instructional approach for the development of second language oral fluency, comprehensibility, and accentedness. *TESOL Quarterly*, *51*, 115-142.

Gluhareva, D., & Prieto, P. (2017). Training with rhythmic beat gestures benefits L2 pronunciation in discourse-demanding situations. *Language Teaching Research*, *21*, 609-631.

Gonzales-Bueno, M., & Quintana-Lara, M. (2011). The teaching of L2 pronunciation through processing instruction. *Applied Language Learning*, *21*, 53-78.

Gooch, R., Saito, K., & Lyster, R. (2016). Effects of recasts and prompts on L2 pronunciation development: Teaching English/ɹ/to Korean adult EFL learners. *System*, *60*, 117-127.

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

Gorsuch, G. J. (2011). Improving speaking fluency for international teaching assistants by increasing input. *TESL-EJ*, *14*, n4.

Hardison, D. M. (2004). Generalization of computer-assisted prosody training: quantitative and qualitative findings. *Language Learning & Technology, 8*, 34-52

Hardison, D. M. (2005). Contextualized computer-based L2 prosody training: Evaluating the effects of discourse context and video input. *CALICO Journal*, *22*, 175-190.

Hazan, V., Sennema, A., Iba, M., & Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication, 47,* 360–378

He, P., & Wasuntarasophit, S. (2015). The effects of video dubbing tasks on reinforcing oral proficiency for Chinese vocational college students. *Asian EFL Journal*, *17*, 106-133.

Henderson, A. (2008). Towards intelligibility: Designing short pronunciation courses for advanced field experts. *ASp-La revue du GERAS*, *53*, 89-110.

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

Hincks, R. (2003). Speech technologies for pronunciation feedback and evaluation. *ReCALL, 15,* 3-20.

Hincks, R., & Edlund, J. (2009). Promoting increased pitch variation in oral presentations with transient visual feedback. *Language Learning & Technology*, *13*, 32-50.

Hirata, Y. (2004). Computer assisted pronunciation training for native English speakers learning Japanese pitch and durational contrasts. *Computer Assisted Language Learning*, *17*, 357-376.

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

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

Kennedy, S., Blanchet, J., & Trofimovich, P. (2014). Learner pronunciation, awareness and instruction in French as a second language. *Foreign Language Annals*, *47*, 79-96.

Kim, J., & Kim, J. (2011). The effectiveness of robot pronunciation training for second language acquisition by children: Segmental and suprasegmental feature analysis approaches. *International Journal of Robots, Education and Art, 1,* 1-17.

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

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

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

Lappin-Fortin, K., & Rye, B. J. (2014). The use of pre-/posttest and self-assessment tools in a French pronunciation course. *Foreign Language Annals*, *47*, 300-320.

Lee, A. H., & Lyster, R. (2017). Can corrective feedback on second language speech perception errors affect production accuracy? *Applied Psycholinguistics*, *38*, 371-393.

Levis, J. M., Sonsaat, S., Link, S., & Barriuso, T. A. (2016). Native and nonnative teachers of L2 pronunciation: Effects on learner performance. *TESOL Quarterly*, *50*, 894-931.

Liakin, D., Cardoso, W., & Liakina, N. (2015). Learning L2 pronunciation with a mobile speech recognizer: French /y/. *CALICO Journal*, *32*, 1-25.

Lima, E. F. (2016). Comprehensibility and liveliness in nonnative student oral presentations before and after training: A mixed methods study. *System*, *63*, 121-133.

Liu, Q., & Fu, Z. (2011). The combined effect of instruction and monitor in improving pronunciation of potential English teachers. *English Language Teaching*, *4*, 164-170.

Lord, G. (2005). (How) can we teach foreign language pronunciation? On the effects of a Spanish phonetics course. *Hispania*, 557-567.

Lord, G. (2008). Podcasting communities and second language pronunciation. *Foreign Language Annals*, *41*, 364-379.

Lord, G. (2010). The combined benefits of instruction and immersion on L2 pronunciation. *Foreign Language Annals, 43,* 488-503.

Martinsen, R., Montgomery, C., & Willardson, V. (2017). The Effectiveness of Video-Based Shadowing and Tracking Pronunciation Exercises for Foreign Language Learners. *Foreign Language Annals*, *50*, 661-680.

Nagamine, T. (2011). Effects of hyper pronunciation training method on Japanese University Students' pronunciation. *Asian EFL Journal Teaching Articles, 53,* 35-50.

Neri, A., Cucchiarini, C., & Strik, H. (2008). The effectiveness of computer-based speech corrective feedback for improving segmental quality in L2 Dutch. *ReCALL*, *20*, 225-243.

Offerman, H. M., & Olson, D. J. (2016). Visual feedback and second language segmental production: The generalizability of pronunciation gains. *System*, *59*, 45-60.

Parlak, Ö. (2010). Does pronunciation instruction promote intelligibility and comprehensibility? *SPLIS News, 7,* 1-5.

Parlak, Ö., & Ziegler, N. (2017). The impact of recasts on the development of primary stress in a synchronous computer-mediated environment. *Studies in Second Language Acquisition*, *39*, 257-285.

Perlmutter, M. (1989). Intelligibility rating of L2 speech pre and post intervention. *Perceptual and Motor Skills, 68,* 515-521.

Rahimi, M. (2017). Second language articulatory training and computer-generated feedback in L2 pronunciation improvement. *ITL-International Journal of Applied Linguistics*, *167*, 190-209.

Rubrecht, B. G. (2004). Teaching/l/and/r/to Japanese EFL Learners: Support for Segmental-Level Pronunciation Instruction. *Asian EFL Journal, 6*, 254-271.

Sadat-Tehrani, N. (2017). Teaching English stress: A case study. *TESOL Journal*, *8*, 943-968.

Saito, K. (2007). The influence of explicit phonetic instruction on pronunciation in EFL settings: The case of English vowels and Japanese learners of English. *Linguistics Journal*, *2*.

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

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

Saito, K. (2013b). Reexamining effects of form-focused instruction on L2 pronunciation development: The role of explicit phonetic information. *Studies in Second Language Acquisition, 35,* 1-29.

Saito, K. (2015). Communicative focus on second language phonetic form: Teaching Japanese learners to perceive and produce English /ɹ/ without explicit instruction. *Applied Psycholinguistics*, *36*, 377-409.

Saito, K., & Lyster, R. (2012). Effects of form-focused instruction and corrective feedback on L2 pronunciation development of/ɹ/by Japanese learners of English. *Language Learning*, *62*, 595-633.

Saito, K., & Lyster, R. (2012). Investigating the pedagogical potential of recasts for L2 vowel acquisition. *TESOL Quarterly*, *46*, 387-398.

Saito, Y., & Saito, K. (2017). Differential effects of instruction on the development of second language comprehensibility, word stress, rhythm, and intonation: The case of inexperienced Japanese EFL learners. *Language Teaching Research*, *21*, 589-608.

Stenson, N., Downing, B., Smith, J., & Smith, K. (1992). The effectiveness of computer-assisted pronunciation training. *CALICO Journal, 9,* 5-19.

Sturm, J. (2013). Explicit phonetics instruction in L2 French: A global analysis of improvement. *System, 41,* 654-662.

Tanner, M., & Landon, M. (2009). The effects of computer-assisted pronunciation readings on ESL learners' use of pausing, stress, intonation, and overall comprehensibility. *Language Learning and Technology, 13,* 51-65.

Thai, C., & Boers, F. (2016). Repeating a monologue under increasing time pressure: Effects on fluency, complexity, and accuracy. *TESOL Quarterly*, *50*, 369–393.

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

Trofimovich, P., Kennedy, S., & Blanchet, J. (2017). Development of second language French oral skills in an instructed setting: A focus on speech ratings. *Canadian Journal of Applied Linguistics/Revue canadienne de linguistique appliquée*, *20*, 32-50.

Trofimovich, P., Lightbown, P. M., Halter, R. H., & Song, H. (2009). Comprehension-based practice: The development of L2 pronunciation in a listening and reading program. *Studies in Second Language Acquisition*, *31*, 609-639.

Tsiartsioni, E. (2010). The effectiveness of pronunciation teaching to Greek state school students. *Advances in Research on Language Acquisition and Teaching,* 429-446.

Underwood, P., & Wallace, M. (2012). The effects of instruction in reduced forms on the performance of low-proficiency EFL university students. *The Asian EFL Journal*, *14*, 1-24.

Weinberg, A., & Knoerr, H. (2003). Learning French pronunciation: Audiocassettes or multimedia? *CALICO Journal*, 315-336.

Yanli, L. (2008). The effectiveness of interactive instruction on the intonation learning of Chinese college learners. *Cross-Cultural Communication, 4,* 90-103.

Supporting Information-B: Results of Coding

Study	Focus of measurement	Scoring method	Task type
de Bot & Mailfert (1982)	Specific	Subjective	Controlled
de Bot (1983)	Specific	Subjective	Controlled
Perlmutter (1898)	Global	Objective	Spontaneous
Stenson et al. (1992)	Specific	Objective	Controlled
Champagne-Muzar & Schneiderman (1993)	Global	Subjective	Controlled
Elliot (1995)	Global, Specific	Subjective	Controlled, Spontaneous
Bradlow et al. (1997)	Specific	Subjective	Controlled
Derwing et al. (1997)	Global	Subjective	Controlled
Ferrier et al. (1999)	Global, Specific	Subjective, Objective	Controlled
Hincks (2003)	Specific	Objective	Controlled, Spontaneous
Weinberg & Knoerr (2003)	Specific	Subjective	Controlled, Spontaneous
Hardison (2004)	Specific	Subjective	Controlled
Hirata (2004)	Specific	Subjective	Controlled
Akita (2005)	Specific	Objective	Controlled
Hardison (2005)	Specific	Subjective	Spontaneous
Hazan (2005)	Specific	Subjective	Controlled
Lambacher (2005)	Specific	Subjective, Objective	Controlled
Lord (2005)	Specific	Objective	Controlled
Couper (2006)	Specific	Objective	Controlled
AbuSeileek (2007)	Specific	Subjective	Controlled
Rubrecht (2007)	Specific	Subjective	Controlled
Saito (2007)	Specific	Objective	Controlled
Henderson (2008)	Specific	Objective	Controlled, Spontaneous
Lord (2008)	Global	Subjective	Controlled
Neri (2008)	Specific	Subjective	Controlled
Yanli (2008)	Specific	Subjective, Objective	Controlled
Hincks & Edlund (2009)	Specific	Objective	Controlled, Spontaneous
Tanner & London (2009)	Global, Specific	Subjective	Controlled, Spontaneous
Trofimovich et al. (2009)	Global, Specific	Subjective, Objective	Controlled
Bueno Alastuey (2010)	Global, Specific	Subjective	Spontaneous
Lord (2010)	Specific	Objective	Controlled
Parlak (2010)	Global	Subjective	Spontaneous
Tsiartsioni (2010)	Specific	Objective	Controlled
Alves & Magro (2011)	Specific	Objective	Controlled
Chen & Goswami (2011)	Specific	Subjective	Controlled

de Jong & Perfetti (2011)	Specific	Objective	Spontaneous
Gonzales-Bueno & Quintana-Lara (2011)	Specific	Objective	Controlled
Gorsuch (2011)	Specific	Subjective, Objective	Controlled, Spontaneous
Kim & Kim (2011)	Specific	Objective	Controlled
Liu & Fu (2011)	Specific	Subjective	Controlled
Nagamine (2011)	Specific	Objective	Controlled
Saito (2011)	Global	Subjective	Controlled, Spontaneous
Thomson (2011)	Specific	Subjective	Controlled
Iverson et al. (2012)	Specific	Subjective	Controlled
Saito & Lyster (2012a)	Specific	Subjective, Objective	Controlled, Spontaneous
Saito & Lyster (2012b)	Specific	Objective	Controlled
Underwood & Wallace (2012)	Specific	Subjective	Spontaneous
Herd (2013)	Specific	Subjective	Controlled
Huensch (2013)	Specific	Subjective	Controlled
Kissling (2013)	Specific	Subjective, Objective	Controlled
Saito (2013a)	Specific	Subjective	Controlled, Spontaneous
Saito (2013b)	Specific	Objective	Controlled, Spontaneous
Sturn (2013)	Specific	Subjective	Controlled
Derwing et al. (2014)	Global, Specific	Subjective, Objective	Controlled, Spontaneous
Kennedy et al. (2014)	Specific	Subjective, Objective	Controlled, Spontaneous
Kissling (2014)	Specific	Subjective	Controlled
Lappin-Fortin (2014)	Specific	Subjective	Controlled
Crowther (2015)	Global	Subjective	Spontaneous
He & Wasuntarasophit (2015)	Global	Subjective	Spontaneous
Liskin et al. (2015)	Specific	Objective	Controlled
Saito (2015)	Specific	Objective	Controlled, Spontaneous
Gooch et al. (2016)	Specific	Subjective	Controlled, Spontaneous
Gordon & Darcy (2016)	Global, Specific	Subjective, Objective	Controlled
Levis et al. (2016)	Global	Subjective	Controlled, Spontaneous
Lima (2016)	Global	Subjective	Spontaneous
Offerman & Olson (2016)	Specific	Objective	Controlled, Spontaneous
Foote & McDonough (2017)	Global	Subjective	Spontaneous
Galante & Thomson (2017)	Global	Subjective	Spontaneous
Gluhareva & Prieto (2017)	Global	Subjective	Spontaneous
Lee & Lyster (2017)	Specific	Subjective	Controlled

Martinsen et al. (2017)	Global	Subjective	Controlled, Spontaneous
Parlak & Ziegler (2017)	Specific	Objective	Spontaneous
Rahimi (2017)	Specific	Objective	Controlled
Sadat-Tehrani (2017)	Specific	Subjective	Controlled
Saito & Saito (2017)	Global, Specific	Subjective, Objective	Controlled
Thai & Boers (2017)	Specific	Objective	Spontaneous
Trofimovich et al. (2017)	Global	Subjective	Controlled, Spontaneous