tesol QUARTERLY
tesol international association

What Characterizes Comprehensible and Native-like Pronunciation Among English-as-a-Second-Language Speakers? Meta-Analyses of Phonological, Rater, and Instructional Factors

KAZUYA SAITO
University College London
London, UK

Abstract

The current study presents two meta-analyses to explore what underlies the assessment and teaching of comprehensible and nativelike pronunciation among English-as-a-Second-Language speakers. In Study 1, listener studies (n = 37) were retrieved examining the influence of segmental, prosodic, and temporal features on listeners' intuitive judgements of comprehensibility and nativelikeness/accentedness as per different listener backgrounds (expert, mixed, L2). In Study 2, training studies (n = 17) were retrieved examining the effects of segmental, prosodic, and temporal-based instruction on ESL learners' pronunciation. The results showed that (a) comprehensibility judgements were related to a range of segmental, prosodic, and temporal features; (b) accentedness judgements were strongly tied to participants' correct pronunciation of consonants and vowels; and (c) instruction led to larger gains in comprehensibility than in nativelikeness. Moderator analyses demonstrated that expert listeners were more reliant on phonological information. Greater effects of instruction on comprehensibility than nativelikeness became clearer, especially when the treatment targeted prosodic accuracy. The findings suggest that ESL practitioners should prioritize suprasegemental practice to help students achieve comprehensible L2 pronunciation. The attainment of nativelike pronunciation, by contrast, may require an exclusive focus on the refinement of segmental accuracy, which is resistant to the influence of instruction.

doi: 10.1002/tesq.3027

Attaining native-like pronunciation has long been considered a pedagogical priority in English-as-a-Second-Language (ESL) classrooms all over the world (e.g., Scales, Wennerstrom, Richard, & Wu, 2006). However, experts in the field of second language (L2) pronunciation have pointed out that the majority of adult L2 speech ends up foreign-accented (e.g., Flege, Munro, & MacKay, 1995). Because of this, the criteria underlying L2 pronunciation assessment and teaching should arguably prioritize communicative success over native-like production (e.g., Derwing & Munro, 2015). The current investigation reports on the results of two meta-analytic studies which examine the factors that underlie listeners' intuitive judgements of comprehensible versus native-like L2 English speech. The findings of these analyses are intended to not only inform L2 pronunciation pedagogy but also to draw tentative conclusions regarding our theoretical understanding of the complex relationship among speakers, listeners, L2 speech assessment, and acquisition.

BACKGROUND

Assessing Second Language Pronunciation

It is well documented that L2 pronunciation is coloured by phonological and phonetic features found in the first language (L1), especially when the onset of learning begins after puberty (Flege et al., 1995). Much of the material for teaching and learning pronunciation is driven by a *nativelikeness* orientation (e.g., Foote et al., 2011 for ESL in Canada), which seeks to reduce or eliminate L1 accent from L2 speech (Tokumoto and Shibata, 2011). The attainment of native-like pronunciation, however, may be limited to individuals with specific cognitive-perceptual abilities, such as phonemic coding (Hu et al., 2013), associative memory (Silbert et al., 2015), and precise auditory processing and acuity (Saito, Kachlicka, Sun, & Tierney, 2020). Furthermore, very few learners are able to reach native-like pronunciation norms, and may only be able to do so if their L1 is linguistically close to the target language (e.g., Dutch learners of English; Bongaerts, van Summeren, Planken, & Schils, 1997; see also Saito, Macmillan, et al., 2020 for Indo-European vs. non Indo-European speakers of L2 English).

In light of these findings, it is important that language teachers are made aware that attaining native-like L2 pronunciation is a difficult task—even if it is an idealized goal. It may also be an unnecessary one, however, considering that much English-medium communication takes place between L2 users themselves. In this setting, foreign accent is a normal and expected characteristic of L2 speech (Pennycook, 2017). On the basis of this argument, a number of scholars have emphasized the importance

of attaining the more *realistic* and *achievable* goals of comprehensibility, intelligibility, and communicative adequacy, as these are what ultimately matter for successful L2 communication (Levis, 2018).

Comprehensibility and Accentedness

The concepts, methodologies, and operationalizations of "realistic" pronunciation goals have widely varied in primary studies (e.g., Munro & Derwing, 2011 for a list of different outcome measures for "intelligibility"; for further discussion, see the Future Directions section in this paper). The current study focuses on two global constructs of L2 pronunciation proficiency: comprehensibility (i.e., ease of understanding) and accentedness (i.e., phonological nativelikeness; for more detailed discussion on different dimensions of L2 pronunciation proficiency, see Saito & Plonsky, 2019). Since Derwing and Munro's seminal work (e.g., Derwing & Munro, 1997; Munro & Derwing, 1995), much attention has been given to contrasting L2 comprehensibility and accentedness. From a methodological perspective, both constructs are measured in the same way—by tapping into listeners' *intuitive* judgements of L2 speech. Upon hearing a sample of speech, raters use a 9-point scale to evaluate how comprehensible (*1 = difficult to understand, 9 = easy to understand*) and accented (*1 = heavily accented, 9 = no accent*) that sample was. In other L2 speech assessment studies, accentedness has also been operationalized as "global foreign accent" (e.g., Riney & Takagi, 1999) and "perceived nativelikeness" (Abrahamsson & Hyltenstam, 2009). In essence, these terms (accentedness, global foreign accent, and nativelikeness) refer to a conceptually similar phenomenon—how closely L2 speech approximates the phonological norm of native speakers. However, the concept of accentedness stands in sharp contrast with that of comprehensibility, which is assumed to index listeners' effort, and by extension ease, of understanding.[1]

This intuitive approach to assessing comprehensible and native-like L2 pronunciation has strong ecological validity, as it is assumed to reflect the instant and impressionistic judgements made by interlocutors during oral communication in real-life contexts (whether communication takes place between L1 and L2 speakers or between L2 and L2 speakers). It also differs from expert assessment, where professional coders are trained to determine global proficiency in accordance with detailed rubrics (see Issacs, Trofimovich, Yu, & Muñoz Chereau, 2015 for a discussion of the relationship between comprehensibility and L2 pronunciation proficiency in IELTS).

[1] Although many studies have adopted a 9-point scale, some used different scalar systems (e.g., 5-point vs. 7-point; for further methodological discussion, see Isaacs & Thomson, 2013).

Findings from several studies conducted by Derwing and Munro have shown that it is common for speech to be judged as highly accented yet remain comprehensible, suggesting that comprehensibility and accentedness are essentially different phenomena (Derwing & Munro, 1997; Munro & Derwing, 1995). Examining what characterizes and distinguishes comprehensibility and accentedness is a crucial initiative with a number of practical implications. Thus far, many scholars have delved into which phonological features are relatively important (or irrelevant) to listener judgements of L2 comprehensibility and accentedness judgements. Such studies enable practitioners to identify discrete sets of pronunciation features that students could practice as a priority in order to improve their global L2 pronunciation proficiency in accordance with their goals (i.e., enhancing comprehensibility vs. nativelikeness; Trofimovich & Isaacs, 2012) and intended interlocutors (e.g., L1 vs. L2 listeners; Saito, Tran et al., 2019).

From a theoretical standpoint, comprehensibility (rather than nativelikeness) is crucial for measuring adult L2 speech development. The Interaction Hypothesis (Gass, 1997; Long, 1996; Mackey, 2012), for instance, posits that language learning takes place precisely when input is made comprehensible during conversational interaction between speakers. A great deal of attention has been directed towards investigating L2 speakers' interlanguage development and ultimate attainment of L2 comprehensibility and accentedness in both naturalistic (e.g., Derwing & Munro, 2013; Saito, 2015) and classroom settings (e.g., Nagle, 2018; Saito & Hanzawa, 2018). These studies generally agree that comprehensibility can continue to improve irrespective of the degree of foreign accentedness as long as the L2 is used regularly on a daily basis (for a similar theoretical account of adult L2 speech learning, see Flege & Bohn, 2020). Contrastingly, the incidence of native-like L2 pronunciation is considerably rare among post-pubertal learners, and determined by factors related to learners' special talent rather than the length, quality, and timing of L2 use (e.g., Saito, Kachlicka, et al., 2020 for auditory sensitivity). Therefore, examining the phonological correlates of comprehensibility and accentedness is an important step towards shedding light on the driving force of two major dimensions of L2 speech learning.

Phonological Correlates of Comprehensibility vs. Accentedness

There is ample evidence that comprehensibility judgements are associated with a range of phonological features, including segmental contrasts with high functional load (English [r]-[l] rather than English [s]-[θ]; Munro & Derwing, 2006; Suzukida & Saito, 2019), word stress

and intonation (Kang, Rubin, & Pickering, 2010; Trofimovich & Isaacs, 2012), speech rate (Munro & Derwing, 2001; Saito, Trofimvich, & Isaacs, 2017), and the frequency, length, and location of pauses (Suzuki & Kormos, 2020). Similarly, accentedness has been associated with a range of segmental errors regardless of the status of functional load (Munro & Derwing, 2006), degree of saliency (Riney & Takagi, 1999), and prosodic accuracy/fluency (Trofimovich & Baker, 2006).

Studies adopting a *longitudinal* perspective have also examined the relative importance of specific phonological features in L2 comprehensibility and accentedness by adopting a pre- and post-test design. These studies typically deliver, analyse, and compare the efficacy of different types of instruction that are related to different features of interest (e.g., segmentals vs. suprasegmentals). Results of primary studies have thus far shown that teaching certain features can make a perceptible impact on comprehensibility and accentedness. Features examined so far include English-specific segmentals (Saito, 2011; Wisniewska & Mora, 2020); word/sentence stress and intonation (Gluhareva & Prieto, 2017; Saito & Saito, 2017); and speech clarity, fluidity, and smoothness (Hamada, 2018 for shadowing; Galante & Thomson, 2017 for drama-based techniques; Tran & Saito, in press for 4/3/2 activity).

Listener Factors

In line with Derwing and Munro's framework, comprehensibility and accentedness can be affected by factors related not only to speakers but also to listeners. In other words, even if two listeners assess the same speech stimuli, their ratings may differ to some degree due to, for example, the quantity and quality of their experience with L2 speech assessment. As reviewed above, much discussion has revolved around how L2 speakers can improve the segmental and suprasegmental qualities of their speech (speaker → listener comprehensibility/accentedness). Though fewer in number, some empirical studies have begun to illustrate how listeners' backgrounds influence their comprehensibility judgements, and how they can adjust the strategies used when listening to accented speech (listener → speaker comprehensibility/accentedness) (for an overview, Derwing & Munro, 2015).

A critical line of research has attempted to identify the factors that influence listeners' judgements of L2 comprehensibility and accentedness. For example, it has been shown that listeners tend to assign more lenient scores when they are familiar with particular foreign accents (Kennedy & Trofimovich, 2008) and topics (Gass & Varonis, 1984), have a linguistics and/or teaching background (Saito, Trofimovich, Isaacs, & Webb, 2016; Saito, Trofimovich, & Isaacs, 2017), and

have multilingual experience (Saito & Shintani, 2016; Shintani, Saito, & Koizumi, 2019). A subset of studies has also highlighted the differences and similarities between L1 and L2 listeners' L2 comprehensibility judgements (e.g., Foote & Trofimovich, 2018; Saito, Tran et al., 2019). It is noteworthy, however, that other studies have failed to find significant effects of listener backgrounds in L2 comprehensibility and accentedness judgements (e.g., Isaacs & Thomson, 2013 for experienced vs. trained listeners; Crowther, Trofimovich, & Isaacs, 2016 for L1 vs. L2 listeners).

Reliability Coefficients

Another factor that remains relatively unexplored is the *reliability* of L2 comprehensibility and accentedness judgements. In high-stakes speaking assessments, professional raters receive hours of special training to rate speech using holistic rubrics. These raters practice and reach agreement rates that typically range between 0.70 and 0.80 (e.g., see Chen et al., 2018 for TOEFL speaking tasks). It is noteworthy that comprehensibility and accentedness judgements are made intuitively by listeners without any training or rating descriptors. Thus, one obvious question concerns whether the judgements of untrained listeners can reach a comparable agreement rate, and whether the strength of agreement varies according to listener backgrounds. Although there is some qualitative research which hints that experienced listeners likely have a clear understanding of their assessment processes compared to naïve listeners (Isaacs & Thomson, 2013), to my knowledge, no synthesis of the research has included degree of consistency as a variable of interest.

According to Plonsky and Derrick (2016), the lack of research on the development, discussion, and provision of guidelines on the reliability of comprehensibility and accentedness judgements problematizes future research, since it considerably clouds the interpretability of study findings (i.e., whether to ascribe any parts of results to variables in concern or to unreliable outcome measures). Furthermore, the comparability of studies remains unclear even if the same methods have been used, which, in turn, negatively impacts the construct validity of meta-analyses on this topic. In the field of applied linguistics, some scholars have proposed rough estimates of acceptable rates of consistency (e.g., $\alpha > .70$ as "moderate" to "substantial"; Brown, 2014). Plonsky and Derrick (2016) surveyed different types of reliability estimates reported in 535 primary studies, finding that the benchmark of satisfactory inter-rater reliability could be relatively high ($\alpha = .92$, interquartile range =.13).

MOTIVATION FOR CURRENT STUDY

Over the past 15 years, the distinction between comprehensible and native-like L2 pronunciation has attracted a great deal of attention from researchers and practitioners alike. To further examine precisely what distinguishes between comprehensibility and accentedness in L2 pronunciation assessment and teaching, there are several topics worthy of further investigation which could have important implications for ESL practitioners all over the world. To synthesize what underlies L2 comprehensibility and accentedness judgements, I will present the results of a meta-analysis on the existing literature. In particular, the paper highlights how L2 comprehensibility and accentedness are differentially tied to (a) speaker factors (i.e., which speech properties affect judgements) and (b) listener factors (i.e., how listener backgrounds influence judgements) (Derwing & Munro, 2015). These are taken up as the main issues in the current study.

Speaker Factors

The first topic relates to the process of L2 comprehensibility and accentedness judgements (i.e., what phonological dimensions underlie listeners' instant, intuitive, and scalar judgements of comprehensibility and accentedness). A clearer understanding of these dimensions could lead to numerous practical implications. For assessment, the findings could reveal whether listener behaviours actually differ between the supposedly distinguishable constructs of L2 speech assessment (i.e., perceiving ease of understanding and nativelikeless). For teaching, the findings could directly inform practitioners about which pronunciation features make the most impact on comprehensibility and accentedness, and how learners can be encouraged to achieve two different goals of L2 pronunciation learning in an efficient and effective manner (enhancing comprehensibility vs. reducing foreign accentedness).

As reviewed earlier, a number of primary studies have focused on a range of pronunciation features that significantly affect L2 comprehensibility and accentedness (for a narrative review, Munro, Derwing, & Thomson, 2015). Accordingly, it is unsurprising that studies directly comparing the phonological correlates of comprehensibility and accentedness within a single study have led to different observations. For example, although Munro and Derwing (1995) found that segmental accuracy was a primary determinant of accentedness, Trofimovich and Issacs (2012) found that prosodic accuracy accounted for the largest amount of variance in both comprehensibility and accentedness.

Intervention studies have produced similarly mixed findings. For example, some studies have demonstrated that both suprasegmental- and segmental-based instruction affect comprehensibility and accentedness, especially when its effectiveness was tested via controlled tasks (e.g., Zhang & Yuan, 2020). Other studies, however, suggest that segmental-based instruction is facilitative of comprehensibility but not accentedness (e.g., Saito, 2011), and that suprasegmental-based instruction likely leads to more gains in comprehensibility (e.g., Gordon & Darcy, 2016). These studies have yet to reach a consensus on which pronunciation features actually matter for the assessment and training of L2 comprehensibility and accentedness.

Listener Factors

A second topic worth clarifying is whether the phonological correlates of comprehensibility and accentedness are subject to the influence of listener background. The aforementioned literature review has brought to light the lack of agreement on this topic. In spite of the supporting evidence, some studies have indicated that the role of listener background may be minor (Munro, Derwing, & Morton, 2006) and/or non-quantifiable (Isaacs & Thomson, 2013). In line with Plonsky and Derrick's (2016) call for the further examination of reliability estimates in applied linguistics research, it is crucial to take a first step towards surveying the inter-rater reliability of listeners' intuitive judgements in accordance with their background.

Previous Meta-Analyses

To my knowledge, there are two published meta-analysis studies concerning L2 pronunciation *teaching*, that is, Lee, Jang, and Plonsky (2015) and Saito and Plonsky (2019).[2] These projects examined how instruction could be facilitative of L2 pronunciation development on

[2] During the final publication process of the current manuscript (March 2021), two similar meta-analysis projects were identified to be either published (Suzuki, Kormos, & Uchihara, 2021) or ongoing (Crowther, forthcoming). Using different screening criteria (including diverse L1-L2 pairings), these projects have looked at different dimensions of global L2 pronunciation proficiency. Whereas Suzuki et al.'s focus lies in the acoustic correlates of perceived fluency (rather than accuracy), Crowther's report aims to provide a comprehensive analysis of comprehensibility, accentedness, and intelligibility. Here I would like to claim that the topic (i.e., what matters for listeners' intuitive reactions to foreign accented speech) will continue to grow as an important research agenda in the field, given that the findings of the meta-analyses (including mine) will help us design and carry out future studies with more rigorous methodologies.

a broader level. L2 pronunciation proficiency was conceptualized/operationalized in many different ways such as overall impressions (comprehensibility, accentedness, and fluency), segmental accuracy (the correct pronunciation of consonants and vowels), prosodic accuracy (the lack and misplacement of word and sentence stress), and fluency (speech rate, pause ratio, repair, and self-repetition ratio). The instruction variable was treated as a monolithic construct without any mention of instructional focus (e.g., comprehensibility vs. accentedness; segmentals vs. prosody vs. fluency). More importantly, none of the meta-analyses analysed L2 pronunciation *assessment*; the current study took a first step towards detangling the relationship among constructs (comprehensibility and accentedness), speech properties (phonological accuracy and fluency), and rater backgrounds (expert vs. novice).

Research Questions

The mixed findings on these two important topics in L2 speech research—that is, speaker and listener factors in L2 comprehensibility and accentedness—could be ascribed to a range of methodological differences (e.g., speakers, elicitation methods, listeners, contexts). By synthesizing the outcomes of each primary research via a meta-analytic approach, the current study aims to provide a more comprehensive picture of the mechanisms underlying listeners' judgements of L2 speech. The following three research questions were formulated:

1. What is the observed inter-rater reliability of intuitive L2 comprehensibility and accentedness judgements?
2. Which pronunciation features do listeners use during their judgements of comprehensible and native-like pronunciation?
3. How does listener background influence the strength of agreement, and the relative weight of segmentals, prosody, and fluency in L2 comprehensibility and accentedness judgements?

In order to detangle the multilayered links among speakers, listeners, and L2 judgements, two different meta-analyses are conducted to approach this topic from two different angles. Study 1 focuses on which pronunciation features (segmentals, prosody, and fluency) listeners attend to while assessing the comprehensibility and accentedness of ESL speech (n = 37 listener studies). Study 2 focuses on the extent to which different types of instruction (segmental, prosodic, and temporal practice) can impact on L2 comprehensibility and accentedness in the most effective and efficient way (n = 17 training studies).

STUDY 1: LISTENER RESEARCH

Study Retrieval and Screening

Focused and Narrow Approach. Following Plonsky and Brown's (2015) emphasis on the importance of defining a meta-analytic domain of interest, the scope of the search was carefully determined in conjunction with the objectives of the study. Although the previously published meta-analyses explored L2 pronunciation *teaching* (Lee et al., 2015; Saito & Plonsky, 2019), the current project concerns the *assessment* of L2 pronunciation proficiency. The search specifically focuses on the *pronunciation* factors that affect listeners' intuitive evaluations of the comprehensibility and accentedness of L2 *English* speech. This focus was adopted in order to provide pedagogical implications tailored to ESL practitioners in particular (teachers, students, and assessors). More importantly, there is evidence that the relative importance of L2 comprehensibility and accentedness greatly varies in accordance with different L1-L2 pairings, resulting in different phonetic features and interlanguage issues (Idemaru, Wei, & Gubbins, 2019). Following Plonsky and Brown's (2015) conceptual framework, the current study could be considered a *focused* meta-analysis in that it only included those studies directly relevant to the aims and context of the study.

Inclusion and Exclusion Criteria

The search procedures and inclusion/exclusion criteria used in the current study were inclusive in nature, featuring a wide range of publication sources, such as papers published in peer-reviewed journals, book chapters, research reports, conference proceedings, and PhD dissertations.

First, the literature search was conducted using a range of tools and sources. These included reference sections of primary studies and online search engines. The search engines were linked to six major library databases (Educational Resources Information Center, Linguistics and Language Behavior Abstracts, PsycINFO, PsycArticles, Web of Science, and ProQuest Dissertations) and one online resource (Google and Google Scholar).

Keywords for the search included *accentedness*, *assessment*, *comprehensibility*, *fluency*, *foreign accent*, *intelligibility*, *nativelikeness*, *oral proficiency*, *pronunciation*, *raters*, *speaking*, and *speech*. The publication year of Munro and Derwing's (1995) original paper was set as the starting point, and February 2020 as the final cut-off point.

Following the notion of an inclusive approach, ancestry searches were conducted on a range of peer-reviewed journals (e.g., *Applied Linguistics, Applied Psycholinguistics; Bilingualism: Language and Cognition;*

Journal of Second Language Pronunciation; Language Learning; Language Teaching Research; Modern Language Journal; and *TESOL Quarterly*); key edited volumes, such as the handbook of English pronunciation (Reed & Levis, 2015), the Routledge handbook of contemporary English pronunciation (Kang, Thomson, & Murphy, 2017), and second language pronunciation assessment (Trofimovich & Isaacs, 2017); major conference proceedings (e.g., International Congress of Phonetic Sciences; Pronunciation in Second Language Learning and Teaching; New Sounds); education reports (e.g., IELTS Research Report Series); and PhD dissertations (included in ProQuest Dissertations).

Second, the decision was made to include only studies examining *segmental, prosodic,* and t*emporal* influences on comprehensibility and accentedness judgements (the main objective of the study). Thus, studies examining listener behaviour only (e.g., Trofimovich & Isaacs, 2011) or the lexicogrammar correlates of comprehensibility and accentedness judgements (e.g., Ruivivar & Collins, 2018) were excluded.

Third, studies needed to provide the necessary information for aggregating reliability statistics (e.g., Cronbach's alpha, interclass correlations) and correlation statistics (correlation coefficients). In this regard, a set of studies were excluded that explored how phonological factors affect L2 comprehensibility and accentedness ratings via descriptive analyses (e.g., McBride, 2015), mean-based analyses (e.g., F and t-tests; Sereno, Lammers, & Jongman, 2016), and variance-based analyses (e.g., multiple regression[3]; O'Brien, 2014).

Finally, since the focus of the current study was on L2 English, $k = 5$ studies focusing on L2 Japanese (Idemaru, Wei, & Gubbins, 2018; Saito & Akiyama, 2017), L2 French (Bergeron & Trofimoivch, 2017; Trofimovich, Kennedy, & Blachet, 2017), and L2 Thai (Wayland, 1997) were excluded. The final dataset comprised 37 empirical studies (26 journal articles, 1 education report, 1 book chapter and 9 PhD dissertations) involving 1022 listeners and 1567 speakers in total. These studies provide the necessary statistical information for the meta-analysis focusing on inter-rater reliability ($n = 31$ studies) and correlation coefficients ($n = 27$ studies) (see Supporting Information).

Coding

To examine which pronunciation features listeners attuned to during L2 their comprehensibility and accentedness judgements, the

[3] Results of multiple regression analyses were excluded, because while models provide R values, they represent a combination of multiple predictors. The current investigation concerns the unique contribution of each predictor (segmentals vs. prosody vs. fluency) to L2 comprehensibility and accentedness.

predictor measures used in the primary studies were coded for three dimensions in accordance with Saito and Plonsky's (2019) framework of L2 pronunciation proficiency: (a) pronouncing consonants and vowels correctly (i.e., segmental accuracy), (b) assigning adequate stress at the word and sentence levels (i.e., prosodic accuracy), and (c) delivering speech at an optimal tempo (i.e., temporal fluency). The corresponding segmental, prosodic, and temporal analysis measures are summarized in Supporting Information.

Given that listener background plays a key role in judgements of L2 comprehensibility and accentedness (Isaacs & Thomson, 2013), the listeners featured in each primary study were coded as either expert or novice. In line with Yan and Ginther (2017), expert listeners were defined as those with previous academic knowledge of linguistics and/or L2 teaching experience, whereas novice listeners were defined as those without such relevant experience. Given the nature of the dataset, two coding groups emerged: expert listeners (n = 12 studies) and mixed listeners (n = 21 studies). The latter category was applied to studies that included both expert and novice listeners. In addition, given that most of communication in English takes place between L2 users, we added another category: L2 listeners (n = 5 studies). Since there were only two studies examining L2 users' accentedness ratings (Crowther et al., 2017; del Río San Román, 2013), the moderator analyses were restricted to the dimension of L2 comprehensibility ratings.

Finally, all reliability estimates were included where available. In line with Plonsky and Derrick's (2016) suggested methodology, all relevant information including interclass correlations and Cronbach's alpha was combined and aggregated. This was because the purpose of the reliability meta-analyses was to provide a *rough* estimate of the distribution and variability of reliability coefficients. To examine the effects of listener backgrounds, inter-rater reliability was for each group of listeners: (a) expert, (b) L2 listeners, and (c) mixed. In all, 37 studies were initially coded by the first author. To check and ascertain the reliability of the coding, a PhD student in applied linguistics was trained using the coding scheme and separately coded approximately 50% of the data (n = 20 studies). There were no coding discrepancies. Thus, the author completed the coding of the remaining studies alone (n = 17 studies).

Results

Average Reliability. Following the analytic procedure recommended in Plonsky and Derrick (2016), the coefficients of reliability (interclass correlations, Cronbach's alpha) were aggregated among a total of 31 empirical studies, where researchers reported the reliability

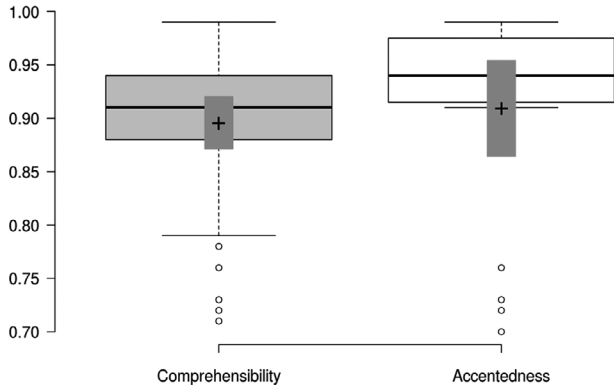

FIGURE 1. Boxplots and 95% Confidence Intervals of the Means (Crosses) for Interrater Agreement for L2 Comprehensibility and Accentedness Judgements

of listeners' comprehensibility and accentedness judgements. As visually plotted in Figure 1, the average inter-rater agreement was relatively high for both comprehensibility and accentedness ($M_{coefficients}$ =.896 for comprehensibility and.909 for accentedness). Descriptive statistics (summarized in Table 1) indicated substantial overlaps in 95% confidence intervals (CI) across the rating dimensions (comprehensibility vs. accentedness) and listener backgrounds, suggesting that listeners generally had strong agreement about the definitions of each dimension.

Effect Size Aggregation. Effect sizes (inverse-variance weighted mean correlation) for the second research question were aggregated using the metafor package (Viechtbauer, 2010) in the R statistical environment (R Core Team, 2018). A random effects model was used that

TABLE 1
Summary of Coefficients of Inter-rater Reliability According to Listener Types

			95% CI	
	k	*M*	Lower	Upper
A. Comprehensibility				
Overall	37	.896	.871	.920
Expert	15	.870	.828	.912
L2	7	.904	.850	.957
Mixed	15	.916	.875	.958
B. Accentedness				
Overall	20	.909	.864	.954
Expert	5	.872	.719	1.024
Mixed	12	.910	.909	1.019

included the main moderator variable. For each primary study, the associations between L2 global ratings (comprehensibility and accentedness) and predictor phonological variables (segmentals, prosody, and fluency) were converted using Fisher's z-transformation ($z = 0.5*\ln(1 + r)/(1 - r)$). Absolute values of the effect sizes were used given the different rating scales used in the studies (*1 = not incomprehensible, 9 = comprehensible* or vice versa) and directionality (e.g., positive or negative) of effects. The z-scores were subsequently transformed back into r values to present the results. Strength of effect size was interpreted using Plonsky and Oswald's (2014) field-specific benchmarks (r =.25,.40, and.60 for small, medium, and large effects, respectively). A within-group Q value (Q_w) was used as a measure of homogeneity for each group's effect sizes (to check the presence of a significant variation in true effect sizes across studies).

To calculate the phonological correlates of L2 comprehensibility and accentedness judgements, a total of 406 effect sizes (from 27 individual studies) were aggregated to produce a weighted mean effect size and 95% CI. A total of n = 274 effect sizes were obtained for comprehensibility from 20 studies, whereas n = 132 were obtained for accentedness from 14 studies. In line with similar meta-analysis projects on L2 pronunciation (Lee et al., 2015; Saito & Plonsky, 2019), an inclusive approach was adopted, allowing one primary study to contribute multiple effect sizes (M = 15.0 effect sizes per study, *range* = 4–25). Given that the studies operationalized the constructs of pronunciation in multiple ways (e.g., fluency as listeners' judgements vs. acoustic analyses of speech rate, pause ratio, and repetition frequency), the decision was made to include as many raw measures as possible (instead of aggregating them). As such, the current meta-analysis was assumed to capture a wide range of methodological variation in primary studies.[4]

As summarized in Table 2, the overall relationship between the phonological predictors and global judgements was moderate-to-strong for both comprehensibility (r =.580, 95% CI [.546,.611], z = 13.427, p <.001) and accentedness (r =.589, 95% CI [.552,.624], z = 10.262 p <.001) (see also Figure 2). As indicated by the overlapping 95% CI values, the listeners' judgements of comprehensibility and accentedness were predicted by phonological accuracy information in particular

[4] It needs to be acknowledged that the inclusive approach may leave the outcomes subject to the influence of a single study. This could be problematic, especially when primary studies feature outliers. According to Tables 3 and 4, CI values did not show much variation (e.g., 0.1–0.2). In addition, the results of Grubbs' tests failed to find any significant outliers in any contexts (p <.05). The results suggest that the dataset well represents how a range of phonological measures are related to L2 comprehensibility and accentedness among the 27 primary studies.

TABLE 2

Summary of Effect Sizes of Phonological Correlates of Comprehensibility and Accentedness Judgements

			95% CI		Homogeneity	
	k	*r*	Lower	Upper	Q_w	*p*
A. Comprehensibility						
Overall	274	.580	.546	.611	1247.055	<.001
Segmentals	44	.640	.561	.707	230.956	<.001
Prosody	122	.587	.536	.633	559.691	<.001
Fluency	108	.546	.489	.598	420.790	<.001
B. Accentedness						
Overall	132	.589	.552	.624	621.997	<.001
Segmentals	24	.792	.735	.838	123.981	<.001
Prosody	64	.587	.533	.636	243.191	<.001
Fluency	44	.456	.379	.528	52.126	<.001

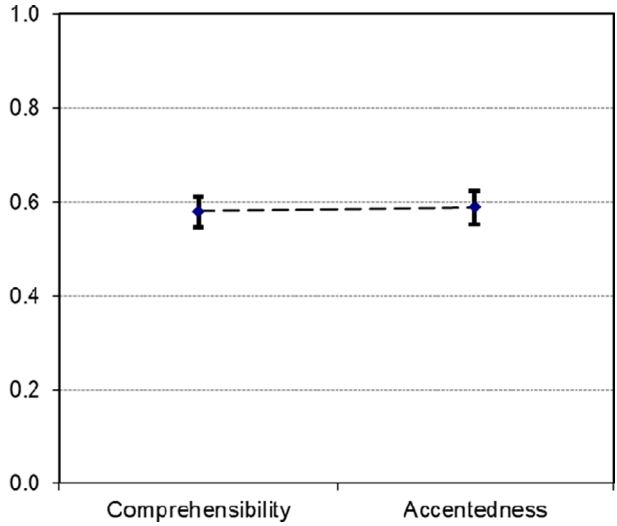

FIGURE 2. 95% Confidence Intervals for the Degree of Phonological Influences on L2 Comprehensibility and Accentedness Judgements

(segmentals and prosody for approximately 30–50% of variances), regardless of the dimension (comprehensibility vs. accentedness).

In terms of the phonological correlates of the judgements, between-group *Q* values were calculated to see whether the strength of the correlation coefficients differed according to the three dimensions (segmentals, prosody, and fluency). Statistical significance was reached for accentedness, $Q(2) = 47.222$, $p < .001$, but not for comprehensibility, $Q(2) = 3.975$, $p = .137$.

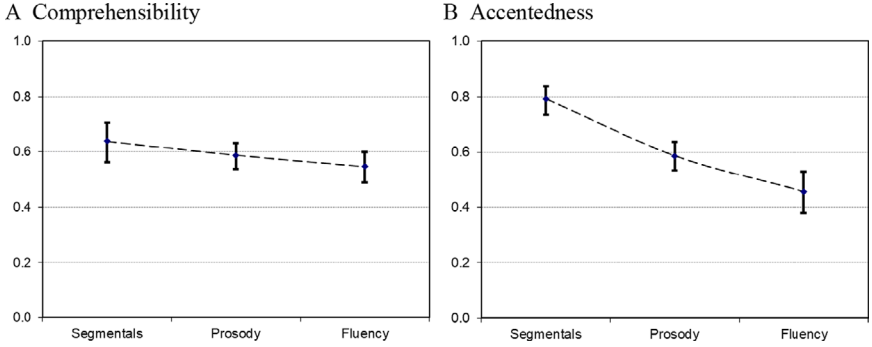

FIGURE 3. Confidence Intervals for the Segmental, Prosodic, and Temporal Correlates of L2 Comprehensibility and Accentedness

As summarized in Table 2, and visually plotted in Figure 3, no significant difference (clear overlaps of 95% CI values) was found among the roles of segmentals, prosody, and fluency in L2 comprehensibility judgements. In contrast, the strength of the correlations between accentedness and the three dimensions of L2 speech (segmentals, prosody, and fluency) were clearly distinguishable at a $p < .05$ level. While judging accentedness, listeners appear to prioritize the three dimensions in the following order: Segmental accuracy ($r = .792$, 95% CI [.735,.838]) > prosodic accuracy ($r = .587$, 95% CI [.533,.636]) > fluency ($r = .456$, 95% CI [.379,.528]).

Listener Backgrounds. The final objective of the statistical analyses was to examine how the moderator variable (i.e., listener background) affected the influence of the phonological features on L2 comprehensibility and accentedness ratings. Descriptive results of the effect sizes according to listener type are summarized in Table 3 and visually plotted in Figure 4. According to the results of Q statistical analyses, the strength of the overall correlations between global and specific pronunciation features differed significantly for comprehensibility among the three different groups of listeners (i.e., expert, mixed vs. L2 listeners), $Q(2) = 25.672$, $p < .001$. The 95% CIs showed that the degree of dependence on phonological information during L2 comprehensibility judgements varied in the following order: experts ($r = .743$, 95% CI [.699,.781]) > L2 listeners ($r = .641$, 95% CI [.582,.693]) > mixed ($r = .442$, 95% CI [.404,.486]). Similarly, the Q analyses showed that the impact of the moderator variable (listener background) reached statistical significance for L2 accentedness judgements, $Q(1) = 16.375$, $p < .001$. Expert listeners' judgements ($r = .758$, 95% CI [.732,.781]) were more strongly associated with phonological information than those of

TABLE 3

Summary of Effect Sizes of Phonological Correlates of Comprehensibility and Accentedness Judgements According to Listener Type

			95% CI		Homogeneity	
	k	*r*	Lower	Upper	Q_w	*p*
A. Comprehensibility						
Overall (expert)	68	.743	.699	.781	336.398	<.001
Segmentals (expert)	14	.786	.697	.851	74.191	<.001
Prosody (expert)	38	.711	.645	.767	169.188	<.001
Fluency (expert)	16	.771	.681	.838	81.429	<.001
Overall (L2)	68	.641	.582	.693	199.388	<.001
Segmentals (L2)	12	.706	.562	.809	17.005	.107
Prosody (L2)	30	.658	.567	.733	104.330	<.001
Fluency (L2)	26	.597	.497	.682	72.076	<.001
Overall (mixed)	138	.446	.404	.486	327.523	<.001
Segmentals (mixed)	18	.442	.321	.548	59.902	<.001
Prosody (mixed)	54	.442	.376	.505	133.185	<.001
Fluency (mixed)	66	.450	.388	.508	131.562	<.001
B. Accentedness						
Overall (expert)	38	.758	.732	.781	218.538	<.001
Segmentals (expert)	7	.919	.896	.937	33.279	<.001
Prosody (expert)	22	.734	.698	.767	54.668	<.001
Fluency (expert)	9	.522	.424	.609	9.115	.332
Overall (mixed)	70	.568	.541	.594	216.941	<.001
Segmentals (mixed)	15	.755	.712	.793	26.430	.028
Prosody (mixed)	32	.559	.519	.596	97.492	<.001
Fluency (mixed)	23	.454	.399	.507	23.220	.389

mixed listeners (r =.568, 95% CI [.541,.594]). Due to the lack of primary studies (n = 2), moderator analysis was not performed for L2 listeners for accentedness.

Lastly, the role of listener background in L2 comprehensibility and accentedness judgements was analysed. As visually plotted in Figure 5, the results showed that for comprehensibility judgements, the listener factor did not seem to impact rating behaviours. All listeners drew equally on the dimensions of accuracy (segmental and prosodic) and fluency, $Q(2)$ = 2.612, p =.271 for expert listeners, $Q(2)$ = 2.006, p =.366 for L2 listeners, and $Q(2)$ = 0.035, p =.982 for mixed.

Listener background played some role in the relationship between phonological factors and the evaluations of L2 accentedness. Overall, both expert and mixed listeners used segmental accuracy as a primary factor during their judgments, $Q(2)$ = 34.681, p <.001 for expert, and $Q(2)$ = 22.998, p <.001. Taking a look at their CI values, expert listeners clearly distinguished between three different dimensions of L2 speech (segmentals [.896,.937] > prosody [.698,.767] > fluency [.424,.609]); however, mixed listeners similarly relied on prosody [.519,.596] and fluency information [.399,.507].

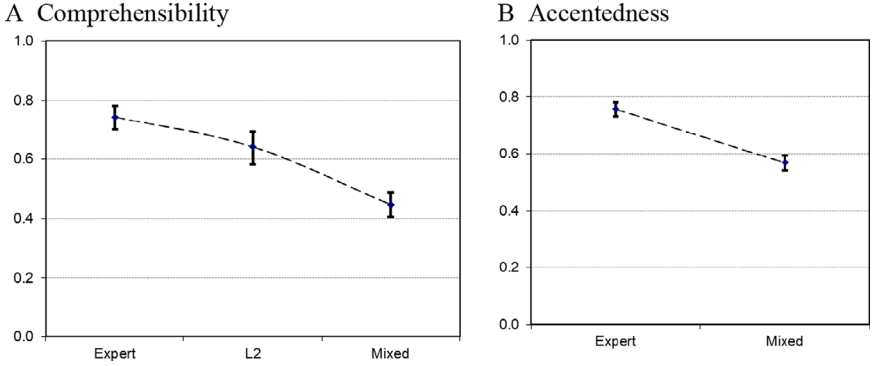

FIGURE 4. 95% Confidence Intervals for the Degree of Phonological Influences on L2 Comprehensibility and Accentedness Judgements Across Differnet Listener Groups

STUDY 2: INTERVENTION RESEARCH

The results of Study 1 suggest that (a) L2 comprehensibility is linked to various phonological features; (b) L2 accentedness is tied to segmental accuracy; and (c) expert listeners rely more on segmental information in their judgements of accentedness in particular. Given that these suggestions were based on cross-sectional research, any discussion of causal relationships needs to be made with caution. To take a different look at the phonological characteristics of L2 comprehensibility and accentedness, scholars have also conducted training studies with a pre- and post-test design. These studies *longitudinally* examine how ESL students develop their pronunciation ability over time when receiving different forms of explicit pronunciation instruction (e.g., segmental-, prosody-, and fluency-based training). Study 2 was designed to meta-analyse the published intervention research so as to shed light on which pronunciation features most impact the *development* of L2 comprehensibility and accentedness.

Lee et al., (2015) and Saito and Plonsly (2019) demonstrated that instruction could facilitate L2 pronunciation learning with small-to-medium effects (d = 0.80, 0.73). However, none of the studies further delved into the extent to which such instructional effectiveness differs when learning gains are assessed for comprehensibility vs. nativelikeness, and the extent to which type of instruction (segmental, prosody, vs. fluency-based training) could maximize the development of comprehensible vs. native-like L2 pronunciation proficiency. The current meta-analysis (Study 2) was designed to correspond to these concerns.

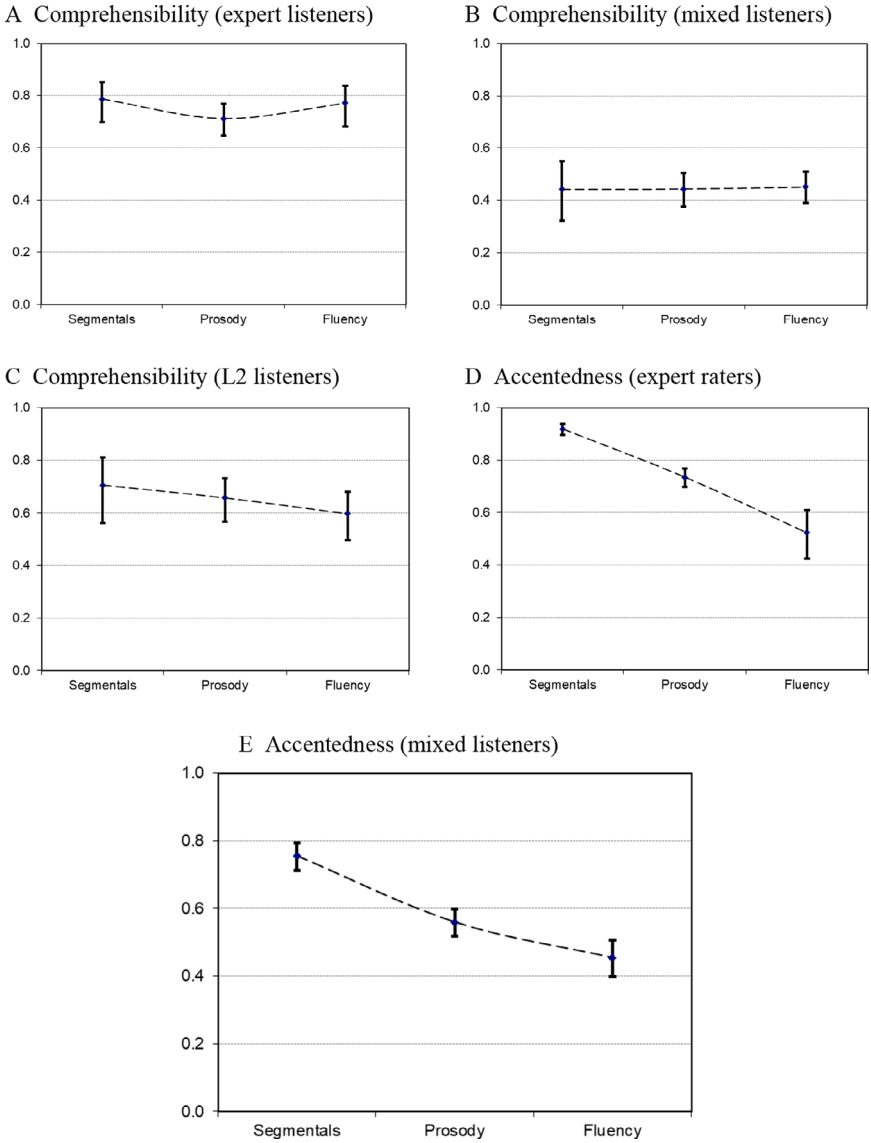

FIGURE 5. 95% Confidence Intervals for the Segmental, Prosodic, and Temporal Correlates of L2 Comprehensibility and Accentedness Across Different Listener Groups

Study Retrieval and Inclusion and Exclusion Criteria

The same search procedures used in Study 1 were adopted here with the addition of the following key words: *instruction*, *teaching*, *pre-*

and post-test, *training*, and *intervention*. The following inclusion criteria which derived from Study 1 were used:

- A wide range of publications (journal articles, book chapter, conference proceedings, and PhD dissertations) were included.
- Comprehensibility and accentedness were used as outcome measures.
- Participants comprised ESL students.

Given that Study 2 relates to intervention studies, the following two new inclusion criteria were also added:

- Explicit pronunciation instruction was provided (for the definition see below).
- Instructional gains (comprehensibility and accentedness) were measured via a pre- and post-test design.

As in Study 1, the date range was between 1995 and February 2020. The final dataset comprised 17 intervention studies involving 290 students (see Supporting Information). They also provided the necessary statistical information for the calculation of Cohen's *d*—that is, *Mean*, *Standard Deviation*, *Standard Error* or/and *t* values. For each study, effect sizes were calculated to index the extent to which participants improved in terms of the comprehensibility and/or nativelikeness of their L2 speech over time (i.e., pre- to post-tests; within-group contrasts). Due to the substantially small number of studies including a control group which did not receive any pronunciation instruction ($n = 5$ out of 17), effect sizes for the between-group contrasts (Experimental vs. Control) were not calculated. A total of 16 peer-reviewed journal articles and one PhD dissertation were included.

Coding

Type of pronunciation instruction was categorized into (a) segmental training, (b) prosody training, and (c) fluency training. Segmental training referred to the provision of explicit instruction on articulatory and perceptual characteristics of vowels and consonants that ESL learners likely have difficulty with (e.g., Saito, 2011 for English [æ, θ, ð w, l, ɹ] for Japanese ESL students). Prosody training referred to the provision of explicit instruction on lexical and sentence stress and intonation (Levis & Levis, 2016 for the use of picture prompted comparison for sentence stress). Fluency training referred to encouraging the memorization, repetition, and/or reading aloud of *already* scripted sentences. The focus of the fluency training lies in clarity, fluidity, and smoothness of speech delivery rather than segmental and prosodic

accuracy. Examples of this kind of training included listening to and repeating what they heard from podcasts (e.g., Foote & McDonough, 2018 for shadowing) and acting as an imaginary character by reciting scripted lines (e.g., Galante & Thomson, 2017 for drama-based techniques). The author and same linguistically trained coder as in Study 1 separately read the 17 intervention studies and coded the primary focus of pronunciation instruction (segmentals vs. prosody vs. fluency). There were no discrepancies in coding.

Results

Effect Size Aggregation. A total of 17 intervention studies produced 28 effect sizes (i.e., Cohen's *d*) for comprehensibility and 20 effect sizes for accentedness. As in Study 1, all effect sizes were submitted to a random effects model using the metafor package (Viechtbauer, 2010) in the R statistical environment (R Core Team, 2018). All indices were interpreted with reference to Plonsky and Oswald's (2014) benchmarks (r = 0.6, 1.0, and 1.4 for small, medium, and large effects, respectively). As summarized in Table 4, and visually plotted in Figure 6, pronunciation training significantly enhanced participants' comprehensibility and accentedness with small effects (d = 0.610, 0.278), as the 95% CI range did not include zero (0.479, 0.740 for comprehensibility; 0.115, 0.440 for accentedness). According to the results of between-group Q tests (Q_b), the difference was statistically significant between comprehensibility and accentedness, $Q(2) = 4.243$, p =.039. This indicates that the impact of instruction was larger for comprehensibility than accentedness.

Descriptive statistics on the effects of segmental-, prosody-, and fluency-based instruction on comprehensibility and accentedness are

TABLE 4
Summary of Effect Sizes of Instructional Effectiveness on Comprehensibility and Accentedness

			95% CI		Homogeneity	
	k	*d*	Lower	Upper	Q_w	*p*
A. Comprehensibility						
Overall	28	.610	.479	.740	53.496	.001
Segmentals	5	.574	.266	.882	14.243	.006
Prosody	14	.566	.378	.754	26.800	.013
Fluency	9	.692	.467	.917	11.679	.166
B. Accentedness						
Overall	20	.278	.115	.440	22.134	.277
Segmentals	4	.118	−.246	.482	0.144	.986
Prosody	10	.214	−.012	.442	5.728	.766
Fluency	6	.497	.197	.798	13.171	.021

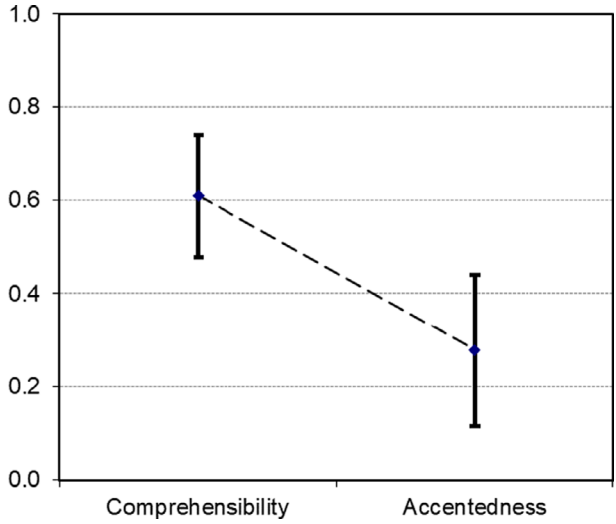

FIGURE 6. 95% Confidence Intervals for the Effects of Instruction on L2 Comprehensibility and Accentedness

TABLE 5
Summary of Effect Sizes of Instructional Effectiveness on Comprehensibility and Accentedness

			95% CI		Homogeneity	
	k	*d*	Lower	Upper	Q_w	*p*
A. Comprehensibility						
Expert	11	.7572	.550	.963	26.475	.003
L2	3	.4703	.113	.827	0.191	.908
Mixed	14	.5242	.332	.715	23.512	.035
B. Accentedness						
Expert	3	.0446	−.461	.551	0.196	.906
Mixed	16	.2928	.120	.465	19.557	.189

summarized in Table 5 and visually plotted in Figure 7. A between-group Q test confirmed that segmental, prosody, and fluency training differentially impacted L2 comprehensibility development and accentedness reduction at a $p < .05$ level, Q_b (5) = 14.454, $p = .013$. According to the results of 95% CI analyses, all the instructional treatments demonstrated CI values above zero, although the lower end of the CI for segmental and prosody training crossed zero for accentedness (−0.246, −0.012, respectively). Thus, the following patterns were suggested: (a) the effectiveness of segmental and prosody training was significant for improving comprehensibility but not accentedness and (b) fluency training improves both comprehensibility and accentedness.

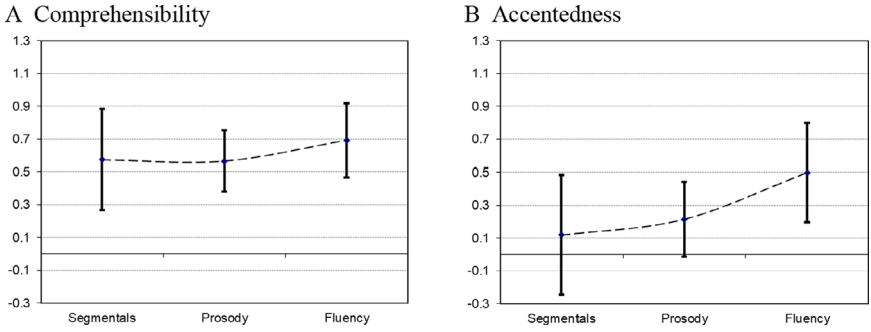

FIGURE 7. Confidence Intervals for the Effects of Segmental, Prosody, and Fluency Training on L2 Comprehensibility and Accentedness

Listener Backgrounds. The final analysis concerns the extent to which listener background (i.e., expert vs. L2 vs. mixed) affects the perceptions of L2 comprehensibility development and accent reduction following pronunciation instruction. The descriptive results are summarized in Table 5 and visually plotted in Figure 8. According to the results of a between-group Q test, the role of listener background was found to be significant, Q_b (4) = 14.502, p =.005. Mixed listeners perceived changes in comprehensibility and accentedness equally, as their 95% CI values were beyond zero. Similarly, L2 listeners' comprehensibility judgements were significant, as the lower end of the 95% CI was beyond zero. However, expert listeners may capture the impact of instruction on comprehensibility but not accentedness.

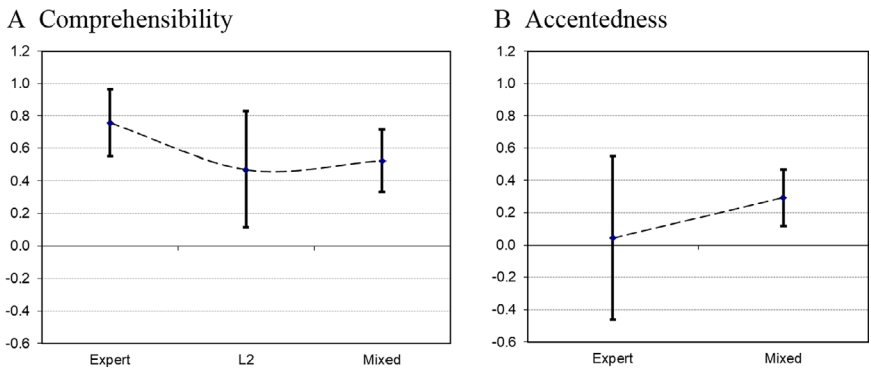

FIGURE 8. 95% Confidence Intervals for the Effects of Instruction on L2 Comprehensibility and Accentedness Across Differnet Listener Groups

DISCUSSION

Global L2 English pronunciation proficiency has been extensively assessed using two inter-related but somewhat independent constructs —comprehensibility (ease of understanding) and accentedness (phonological nativelikeness). These constructs are typically operationalized through listeners' intuitive judgements. In Munro and Derwing's seminal work, degree of comprehensibility and accentedness was assumed to be determined by both the phonological properties of speech (e.g., accuracy and fluency errors; linguistic factors) and by listener background (e.g., amount of prior ESL/EFL teaching, linguistics training; other listener factors). The current study sought to examine these assumptions by meta-analysing the comprehensibility judgements of L2 English speech in 37 listener studies and 17 intervention studies. The analysis generated insightful findings as to the product and process of L2 comprehensibility and accentedness judgements in response to the three research questions.

R1: Reliability of Comprehensibility and Accentedness Judgements

According to the results of the reliability analysis, relatively strong inter-rater agreement was found for both comprehensibility and accentedness judgements (.896,.909), regardless of listener background (expert vs. L2 vs. mixed). The findings suggest that listeners have similar intuitions of which L2 English pronunciation forms are comprehensible and native-like.

R2: Phonological Correlates of Comprehensible and Native-like Pronunciation

In terms of listener behaviours *during* L2 speech assessments, the results showed that approximately 30–50% of the variance in comprehensibility and accentedness judgements could be accounted for by phonological factors. According to Plonsky and Oswald's (2014) benchmarks, these are relatively large effects. The results suggest that listeners rely heavily on phonological accuracy and fluency during their instant and intuitive judgements of L2 speech. However, what distinguishes comprehensibility and accentedness seems to be the type of phonological information listeners actually use when rating. Whereas L2 comprehensibility ratings were equally associated with the

dimensions included in the analysis, L2 accentedness ratings were strongly linked to segmental accuracy in particular.

The meta-analysis of intervention studies further revealed how listeners' perception of comprehensible and native-like pronunciation changes when judging the speech of ESL students who had received segmental, prosody, and/or fluency training. The results showed that the listeners' judgements of comprehensibility were equally associated with the type of instruction received (e.g., segmental, prosody, and fluency training). Comparatively, improvement in accentedness did not appear to be perceptible even if students received segmental and prosodic accuracy training through explicit instruction and/or practice. Given that accentedness proxies, the relatively difficult aspects of L2 speech, that is, the degree of phonological accuracy (see the results of Study 1), it is reasonable to assume that accentedness is resistant to change even when segmental and prosody-focused instruction is provided.

Given that the effectiveness of pronunciation teaching was found to be "small-to-medium" (e.g., Saito & Plonsky, 2019 for d =.73), the current study further demonstrated that such instructional effectiveness may vary according to different foci of assessment and training. Pronunciation teaching makes a "small-to-medium" difference when the focus of assessment highlights the comprehensibility of pronunciation (d =.61); and when the training focuses on fluency (d =.69). However, the effectiveness of training remains small or unclear if it is assessed for nativelikeness (d <.27) and it targets the acquisition of segmental accuracy in particular (d =.12). On the whole, Studies 1 and 2 support the *robustness*, *reliability*, and *replicability* of the findings (i.e., the phonological correlates of comprehensible and native-like pronunciation) across two different types of investigations (cross-sectional vs. longitudinal).

R3: Roles of Listener Factors in Assessment and Teaching of Pronunciation

Finally, the findings showed that there is some differential effect of listener background on L2 comprehensibility and accentedness judgements. In Study 1 (listener studies), expert listeners seemed to rely more on phonological information (segmental accuracy in particular) than novice and mixed listeners, especially when assessing accentedness. The results of Study 2 extend this finding by suggesting that expert listeners consider the impact of instruction to be minor when assessing the accentedness of speakers who had received segmental and prosodic accuracy training. Contrastingly, such listener effects were not clearly observed in any contexts of L2 comprehensibility judgements.

The results imply two possibilities regarding the consequences of using different types of listeners (i.e., mixed listeners) in research on L2 comprehensibility and accentedness. For those who show less reliance on phonological information, non-phonological factors, such as vocabulary (Appel, Trofimovich, Saito, Isaacs, & Webb, 2019), grammar (Ruivivar & Collins, 2018), collocational (Saito, 2020), and discourse knowledge (Trofimovich & Isaacs, 2012) could have been responsible for explaining the remaining variance in ratings (especially when it comes to L2 comprehensibility; see Saito, Trofimovich, Isaacs, & Webb, 2016). Alternatively, it may be that the mechanisms underlying the L2 comprehensibility and accentedness judgements of mixed listeners are inconsistent because of their intricate, multi-layered backgrounds, and potentially random rating behaviour. In other words, mixed listeners may use different types of strategies to arrive at their comprehensibility and accentedness scores (see Nagle, Trofimovich, & Bergeron, 2019 for the use of Idiodynamic Software). This would be a fruitful area of inquiry for future studies (cf. Magne et al., 2019 for quantitative *and* qualitative analysis of rater behaviours during L2 speech assessments).

Revising Framework of L2 Pronunciation Assessment, Teaching, and Development

The findings of the study provide crucial implications for theory building in L2 pronunciation assessment, teaching, and development (as summarized in Table 6). First, they support Derwing and Munro's (2015) listener and speaker model of L2 comprehensibility and accentedness. Two global constructs of L2 pronunciation proficiency—comprehensibility and accentedness—are readily distinguishable as listeners pay equal attention to various areas of phonological information for the former (segmentals = prosody = fluency) and prioritize segmental accuracy for the latter (segmentals > prosody > fluency). Second, listener effects are more clearly observed in accentedness than comprehensibility because expert listeners are more likely sensitive to the nativelikeness of segmental accuracy. The relationship between listener factors and comprehensibility may need to be examined beyond the focus of the current meta-analysis (i.e., phonological dimensions).[5] Third, as conceptualized by the Interaction Hypothesis (e.g., Mackey,

[5] Note that the role of listener background has been indeed identified in lexicogrammar (rather than pronunciation) aspects of L2 speech assessment such that more linguistically trained and experienced raters can better decode/understand what L2 speakers intend to say regardless of phonological non-nativelikeness (e.g., Saito et al., 2016).

TABLE 6
Summary of Framework for L2 Pronunciation Assessment, Teaching, and Development

Global constructs	Phonological correlates	Listener factors	Teaching and development
Comprehensible pronunciation	Segmentals = prosody = fluency	Neutral (listener effects *probably* being more observable in lexicogrammar dimensions)	Amenable to change (due to the learnability in prosody and fluency)
Native-like pronunciation	Segmentals > prosody > fluency	Strong (experts being stricter due to their greater sensitivity to segmental nativelikeness)	Resistant to change (due to the difficulty in segmentals)

2012) and shown in empirical research (e.g., Derwing & Munro, 2013; Saito, 2015), improvement tends to occur in the comprehensibility rather than nativelikeness dimensions of language. This asymmetricity can be explained by the finding that comprehensibility improves as a collective effort of segmental, prosody, and fluency development; and that accentedness is resistant to change because its main component—segmental accuracy—is subject to gradual, extensive, and individually different learning patterns (for longitudinal evidence, Saito, Suzuki, Oyama, & Akiyama, 2020; for more detailed accounts of this topic, see Flege & Bohn, 2020).

Implications for Practitioners

The different phonological correlates of L2 comprehensibility and accentedness identified here have several implications for pronunciation teaching and learning. On the one hand, given that many L2 learners are concerned with sounding native-like, teachers should focus on improving their segmental accuracy—the primary correlate of accentedness (Study 1). However, it is important to remind learners of the mounting evidence that few adult L2 learners can become native-like in their pronunciation (e.g., Flege et al., 1995), as well as to make them aware that accentedness is likely to remain unchanged even after instruction (Study 2). This is because the refinement of L2 segmental accuracy is a slow, gradual, and extensive process, especially beyond the initial stages of learning (Flege et al., 1995; Saito, 2015). Furthermore, it may be subject to the influence of myriad individual differences including motivation (e.g., Moyer, 1999 for professional orientation and commitment; Nagle, 2018 for strong visions of future images), perceptual acuity (e.g., Saito, Kachlicka, et al., 2020), and

cognitive functioning (Darcy, Park, & Yang, 2015 for working memory).

On the other hand, teachers should introduce a range of practice activities which help learners improve the various dimensions of their L2 proficiency in a balanced manner in order to help their students improve the comprehensibility of their speech. The focus of such activities can include segmentals (e.g., Munro & Derwing, 2006 for segmental contrasts with high functional load), prosody (Couper, 2006 for word stress; Saito & Saito, 2017 for intonation), and fluency (Suzuki, 2020; Thai & Boers, 2016; Tran & Saito, in press for timed repetition). In naturalistic L2 speech learning, there is ample cross-sectional and longitudinal evidence that (a) learners can quickly improve the fluency of their speech shortly after starting immersion and (b) L2 prosody will steadily develop as long as learners have access to ample interaction and immersion experience (Trofimovich & Baker, 2006). In fact, the current meta-analysis suggests that the impact of instruction can be clearly observed when speech is assessed in terms of comprehensibility (rather than accentedness).

Implications for Researchers

The current investigation took a first step towards meta-analysing the factors which are most relevant to the assessment and teaching of L2 pronunciation. Specifically, the studies focused on how different phonological dimensions of speech affect judgements of comprehensibility and accentedness, with listeners' backgrounds (expert vs. L2 vs. mixed) as a moderator variable. In light of the significance of the findings, and to provide implications for researchers in particular, I would like to end this paper by proposing the following topics worthy of future meta-analyses.

Intelligibility. While comprehensibility indexes listeners' ease of understanding, there is a consensus that what is ultimately important for communicative success is *intelligibility*, that is, interlocutors' actual understanding of intended message. Although intelligibility is well researched in the field of L2 pronunciation, scholars have continued to debate how to best measure the construct. A diverse range of methods have been used, including transcription, comprehension questions, scaler ratings, and reaction time instruments (for reviews on methodological fuzziness in L2 intelligibility research, see Isaacs, 2008; Kang, Thomson, & Moran, 2018; Munro & Derwing, 2011). Additionally, the existing literature has exclusively relied on audio information as a main source of understanding, although some studies have begun

to examine whether, to what degree, and how audio *and* visual information differentially impact L2 intelligibility (e.g., Drijvers & Özyürek, 2019; Wheeler & Saito, forthcoming).

In a broad sense, comprehensibility is conceptually similar to intelligibility (relative to accentedness), as it was originally referred to as "native speakers' perception of intelligibility" (Derwing & Munro, 1997, p. 2). On a narrower level, comprehensibility is methodologically distinguishable from intelligibility, as it taps into the *actual effort* made to understand (measured via scaler ratings), as opposed to the actual outcome of understanding (assessed via a wide range of measures, such as transcription and comprehension questions). While I make a strong call for future meta-analysis studies to further pursue the mechanisms underlying comprehensibility *and* intelligibility, it also needs to be emphasized that the latter construct should be surveyed with much caution. It may be advisable to wait for more empirically robust methods to be established and for more primary studies using them to be published.

Task. One topic that future meta-analysis studies should explore concerns the conditions of different speaking tasks. Crowther and his colleagues have begun to demonstrate that the phonological correlates of L2 comprehensibility and accentedness may vary according to task structure (Crowther, Trofimovich, Issacs, & Saito, 2015 for simple vs. complex) and formality (Crowther, Trofimovich, Saito, & Issacs, 2018 for academic vs. non-academic). Another distinction relates to controlled vs. spontaneous tasks (Saito & Plonsky, 2019) and structured vs. unstructured (Saito & Liu, 2021). With a sufficient number of primary studies for a robust moderator analysis, future studies can examine the effects of task structure on L2 comprehensibility and accentedness as per any existing task frameworks in the task-based language learning literature (e.g., Robinson, 2011).

Fluency. One interesting finding of the current meta-analysis is that fluency training appears to be equally important for the assessment and development of comprehensible and native-like speech. The existing literature suggests that fluency factors (speech rate, pause frequency) explain a large degree of variance in listeners' judgements of comprehensibility (Suzuki & Kormos, 2020) and accentedness (Trofimovich & Baker, 2006), and that quick, perceptible improvement can be observed in the fluency (rather than accuracy) dimension of L2 speech in both naturalistic and classroom settings (Mora & Valls-Ferrer, 2012; Saito & Hanzawa, 2018). The argument here echoes a growing amount of theoretical discussion that fluency serves as a crucial component of speaking proficiency (see Foster, 2020 for a

comprehensive overview). It would be intriguing for future studies to elaborate on more research-based approaches to fluency instruction, measure its impact on both comprehensibility and accentedness (cf. Suzuki, 2020; Thai & Boers, 2016; Tran & Saito, in press), and promote practitioners' awareness towards the relative importance of fluency (over accuracy) as a component of speaking proficiency (Tavakoli, 2020).

ACKNOWLEDGMENTS

I am grateful to Yui Suzukida for her assistance for data. I would like to thank three anonymous TESOL Quarterly reviewers for their helpful input, feedback and advice at every stage of the manuscript writing and revising processes. The project is funded by Leverhulme Trust Research Grant (RPG-2019-039).

THE AUTHOR

Kazuya Saito is an Associate Professor in Applied Linguistics at University College London, UK. His research interests include how second language learners develop various dimensions of their speech in naturalistic settings; and how instruction can help optimize such learning processes in classroom contexts.

REFERENCES

Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning*, *59*(2), 249–306. https://doi.org/10.1111/j.1467-9922.2009.00507.x

Appel, R., Trofimovich, P., Saito, K., Isaacs, T., & Webb, S. (2019). Lexical aspects of comprehensibility and nativeness from the perspective of native–speaking English raters. *ITL–International Journal of Applied Linguistics*, *170*(1), 24–52. https://doi.org/10.1075/itl.17026.app

Bergeron, A., & Trofimovich, P. (2017). Linguistic dimensions of accentedness and comprehensibility: Exploring task and listener effects in second language French. *Foreign Language Annals*, *50*(3), 547–566. https://doi.org/10.1111/flan.12285

Bongaerts, T., van Summeren, C., Planken, B., & Schils, E. (1997). Age and ultimate attainment in the pronunciation of a foreign language. *Studies in Second Language Acquisition*, 447–465, https://doi.org/10.1017/S0272263197004026

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

Crowther, D. (forthcoming). Global dimensions of second language pronunciation: A meta–analysis of accentedness, comprehensibility, and intelligibility research.

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2015). Does a speaking task affect second language comprehensibility? *The Modern Language Journal*, *99*(1), 80–95. https://doi.org/10.1111/modl.12185

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2018). Linguistic dimensions of L2 accentedness and comprehensibility vary across speaking tasks. *Studies in Second Language Acquisition*, *40*(2), 443–457. https://doi.org/10.1017/S027226311700016X

Darcy, I., Park, H., & Yang, C. L. (2015). Individual differences in L2 acquisition of English phonology: The relation between cognitive abilities and phonological processing. *Learning and Individual Differences*, *40*, 63–72. https://doi.org/10.1016/j.lindif.2015.04.005

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *19*(1), 1–16. https://doi.org/10.1017/S0272263197001010

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence–based perspectives for L2 teaching and research* (Vol. *42*). John Benjamins Publishing Company.

Drijvers, L., & Özyürek, A. (2019). Non–native listeners benefit less from gestures and visible speech than native listeners during degraded speech comprehension. *Language and Speech*, 0023830919831311.

Foote, J. A., Holtby, A. K., & Derwing, T. M. (2011). Survey of the teaching of pronunciation in adult ESL programs in Canada, 2010. *TESL Canada Journal*, 1–22. https://doi.org/10.18806/tesl.v29i1.1086

Foote, J. A., & Trofimovich, P. (2018). Is it because of my language background? A study of language background influence on comprehensibility judgments. *Canadian Modern Language Review*, *74*(2), 253–278. https://doi.org/10.3138/cmlr.2017-0011

Hu, X., Ackermann, H., Martin, J. A., Erb, M., Winkler, S., & Reiterer, S. M. (2013). Language aptitude for pronunciation in advanced second language (L2) learners: Behavioural predictors and neural substrates. *Brain and Language*, *127*, 366–376. https://doi.org/10.1016/j.bandl.2012.11.006

Idemaru, K., Wei, P., & Gubbins, L. (2019). Acoustic sources of accent in second language Japanese speech. *Language and Speech*, *62*(2), 333–357. https://doi.org/10.1177/0023830918773118

Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non–native English–speaking graduate students. *Canadian Modern Language Review*, *64*(4), 555–580. https://doi.org/10.3138/cmlr.64.4.555

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*(2), 135–159. https://doi.org/10.1080/15434303.2013.769545

Isaacs, T., Trofimovich, P., Yu, G., & Chereau, B. M. (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*.

Kang, O., Rubin, D. O. N., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal*, *94*(4), 554–566. https://doi.org/10.1111/j.1540-4781.2010.01091.x

Kang, O., Thomson, R., & Murphy, J. (Eds.). (2017). *The Routledge handbook of contemporary English pronunciation*. London: Routledge, https://www.routledge.com/The-Routledge-Handbook-of-Contemporary-English-Pronunciation/Kang-Thomson-Murphy/p/book/9781138856882.

Kang O., Thomson R. I., Moran M. (2018). Empirical Approaches to Measuring the Intelligibility of Different Varieties of English in Predicting Listener Comprehension. *Language Learning*, *68*, (1), 115–146. http://dx.doi.org/10.1111/lang.12270.

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review*, *64*(3), 459–489. https://doi.org/10.3138/cmlr.64.3.459

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta–analysis. *Applied Linguistics*, *36*(3), 345–366. https://doi.org/10.1093/applin/amu040

Levis, J. M. (2018). *Intelligibility, oral communication, and the teaching of pronunciation*. Cambridge University Press.

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*(4), 894–931. https://doi.org/10.1002/tesq.272

Magne, V., Suzuki, S., Suzukida, Y., Ilkan, M., Tran, M., & Saito, K. (2019). Exploring the dynamic nature of second language listeners' perceived fluency: A mixed-methods approach. *TESOL Quarterly*, *53*, 1139–1150. https://doi.org/10.1002/tesq.528

McBride, K. (2015). Which features of Spanish learners' pronunciation most impact listener evaluations? *Hispania*, 14–30.

Moyer, A. (1999). Ultimate attainment in L2 phonology: The critical factors of age motivation and instruction. *Studies in Second Language Acquisition*, *21*, 81–108. https://doi.org/10.1017/S0272263199001035

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45* (1), 73–97. https://doi.org/10.1111/0023-8333.49.s1.8

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech the role of speaking rate. *Studies in Second Language Acquisition*, *23*(4), 451–468. https://doi.org/10.1017/S0272263106060049

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, *34*(4), 520–531. https://doi.org/10.1016/j.system.2006.09.004

Munro, M. J., & Derwing, T. M. (2011). The foundations of accent and intelligibility in pronunciation research. *Language Teaching*, *44*(3), 316–327. https://doi.org/10.1017/S0261444811000103

Nagle, C. (2018). Motivation, comprehensibility, and accentedness in L2 Spanish: Investigating motivation as a time-varying predictor of pronunciation development. *The Modern Language Journal*, *102*, 199–217. https://doi.org/10.1111/modl.12461

Nagle, C., Trofimovich, P., & Bergeron, A. (2019). Toward a dynamic view of second language comprehensibility. *Studies in Second Language Acquisition*. https://doi.org/10.1017/S0272263119000044

Pennycook, A. (2017). *The cultural politics of English as an international language*. Taylor & Francis.

Plonsky, L., & Brown, D. (2015). Domain definition and search techniques in meta–analyses of L2 research (Or why 18 meta–analyses of feedback have different results). *Second Language Research*, *31*(2), 267–278. https://doi.org/10.1177/0267658314536436

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal*, *100*(2), 538–553. https://doi.org/10.1111/modl.12335

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

Riney, T. J., & Takagi, N. (1999). Global foreign accent and voice onset time among Japanese EFL speakers. *Language Learning*, *49*(2), 275–302. https://doi.org/10.1111/0023-8333.00089

Del Río San Román, C. D. (2013) *Perceived foreign accent and comprehensibility in the oral production of adolescent learners of English: Study abroad vs. at home learning contexts* (Doctoral dissertation, Universitat Pompeu Fabra).

Ruivivar, J., & Collins, L. (2018). The effects of foreign accent on perceptions of nonstandard grammar: A pilot study. *TESOL Quarterly*, *52*(1), 187–198. https://doi.org/10.1002/tesq.374

Saito, K. (2015). Experience effects on the development of late second language learners' oral proficiency. *Language Learning*, *65*, 563–595. https://doi.org/10.1111/lang.12120

Saito, K. (2020). Multi- or single-word units? The role of collocation use in comprehensible and contextually appropriate second language speech. *Language Learning*, *70*, 548–588. https://doi.org/10.1111/lang.12387

Saito, K., & Akiyama, Y. (2017). Linguistic correlates of comprehensibility in second language Japanese speech. *Journal of Second Language Pronunciation*, *3*, 199–218. https://doi.org/10.1075/jslp.3.2.02sai

Saito, K., & Hanzawa, K. (2018). The role of input in second language oral ability development in foreign language classrooms: A longitudinal study. *Language Teaching Research*, *22*, 398–417. https://doi.org/10.1177/1362168816679030

Saito, K., Kachlicka, M., Sun, H., & Tierney, A. (2020). Domain-general auditory processing as an anchor of post-pubertal second language pronunciation learning: Behavioural and neurophysiological investigations of perceptual acuity, age, experience, development, and attainment. *Journal of Memory and Language*, *115*, 104168. https://doi.org/10.1016/j.jml.2020.104168

Saito, K., & Liu, Y. (2021). Roles of collocation in L2 oral proficiency revisited: Different tasks, L1 vs. L2 raters, and cross-sectional vs. longitudinal analyses. *Second Language Research*. https://doi.org/10.1177/0267658320988055

Saito, K., Macmillan, K., Tran, M., Suzukida, Y., Sun, H., Magne, V., … Murakami, A. (2020). Developing, analyzing and sharing multivariate datasets: Individual differences in L2 learning revisited. *Annual Review of Applied Linguistics*, *40*, 9–25. https://doi.org/10.1017/S0267190520000045[Opens in a new window]

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*(3), 652–708. https://doi.org/10.1111/lang.12345

Saito, K., & Shintani, N. (2016). Do native speakers of North American and Singapore English differentially perceive second language comprehensibility? *TESOL Quarterly*, *50*, 421–446. https://doi.org/10.1002/tesq.234

Saito, K., Suzuki, S., & Oyama, T., & Akiyama, Y. (2020). How does longitudinal interaction differentially promote experienced vs. inexperienced learners' L2 speech learning? *Second Language Research*. https://doi.org/10.1177/0267658319884981

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. https://doi.org/10.1093/applin/amv047

Saito, K., Trofimovich., P., Isaacs, T., & Webb., S., (2016). Re–examining phonological and lexical correlates of second Language comprehensibility: The role of rater experience. In T. Isaacs & P. Trofimovich (Eds.), *Interfaces in second language pronunciation assessment: Interdisciplinary perspectives* (pp. 141–156). Bristol, UK: Multilingual Matters.

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. https://doi.org/10.1177/1362168816643111

Sereno, J., Lammers, L., & Jongman, A. (2016). The relative contribution of segments and intonation to the perception of foreign–accented speech. *Applied Psycholinguistics*, *37*(2), 303–322. https://doi.org/10.1017/S0142716414000575 [Opens in a new window]

Shintani, N., Saito, K., & Koizumi, R. (2019). The relationship between multilingual raters' language background and their perceptions of accentedness and comprehensibility of second language speech. *International Journal of Bilingual Education and Bilingualism*, *22*(7), 849–869. https://doi.org/10.1080/13670050.2017.1320967

Silbert, N. H., Smith, B. K., Jackson, S. R., Campbell, S. G., Hughes, M. M., & Tare, M. (2015). Non–native phonemic discrimination, phonological short term memory, and word learning. *Journal of Phonetics*, *50*, 99–119. https://doi.org/10.1016/j.wocn.2015.03.001

Suzuki, S., & Kormos, J. (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*, *42*(1), 143–167. https://doi.org/10.1017/S0272263119000421

Suzuki, S., Kormos, J., & Uchihara, T.(2021). To what extent utterance fluency can predict perceived fluency? A meta–analysis of correlational studies. *Modern Language Journal*.

Suzuki, Y. (2020). Optimizing fluency training for speaking skills transfer: Comparing the effects of blocked and interleaved task repetition. *Language Learning*. https://doi.org/10.1111/lang.12433

Suzukida, Y., & Saito, K. (2019). Which segmental features matter for successful L2 comprehensibility? Revisiting and generalizing the pedagogical value of the functional load principle. *Language Teaching Research*. https://doi.org/10.1177/1362168819858246

Thai, C., & Boers, F. (2016). Repeating a monologue under increasing time pressure: Effects on fluency, complexity, and accuracy. *Tesol Quarterly*, *50*(2), 369–393. https://doi.org/10.1002/tesq.232

Tokumoto, M., & Shibata, M. (2011). Asian varieties of English: Attitudes towards pronunciation. *World Englishes*, *30*(3), 392–408. https://doi.org/10.1111/j.1467-971X.2011.01710.x

Tran, M., & Saito, K. (in press). Effects of the 4/3/2 activity revisited: Extending Boers (2014) and Thai & Boers (2016). *Language Teaching Research*.

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), 1–30. https://doi.org/10.1017/S0272263106060013

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, *15*(4), 905–916. https://doi.org/10.1017/S1366728912000168

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*(2), 32–50. https://doi.org/10.7202/1042675

Viechtbauer, W. (2010). Conducting meta–analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Wayland, R. (1997). Non–native production of Thai: Acoustic measurements and accentedness ratings. *Applied Linguistics*, *18*(3), 345–373. https://doi.org/10.1093/applin/18.3.345

Wheeler, P., & Saito, K. (forthcoming). Second language speech intelligibility revisited: Differential roles of phonological accuracy, visual speech, and iconic gesture.

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

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

Supplementary Material