LANGUAGE LEARNING
A Journal of Research in Language Studies
Language Learning ISSN 0023-8333

Check for updates

EMPIRICAL STUDY

Second Language Sentence Stress Assignment: Self- and Other-Assessment

Cesar Teló he/him [a,b] Hanna Kivistö de Souza [c,d] Mary Grantham O'Brien she/her [e,a] and Angélica Carlet [f]

[a]University of Calgary [b]Concordia University [c]Federal University of Santa Catarina [d]University of Turku [e]Simon Fraser University [f]Charles Darwin University

Research on second language (L2) pronunciation self-assessment reports a general misalignment between self- and other-assessment. This has been attributed to the object of self-assessment, the self-assessment task, the measures to which self-assessment is compared, and speakers' characteristics. Here, we examined self-assessment of a discrete phonological feature—sentence stress—by L2 English speakers as compared to the assessment of first language English listeners through a timed, forced-choice judgment task with low-pass filtered stimuli, which contained only suprasegmental cues. Additionally, we explored how individual differences among speakers predict self-assessment. Speakers generally overestimated their accuracy in sentence stress

CRediT author statement–**Cesar Teló**: conceptualization; methodology; investigation; formal analysis; software; writing–original draft preparation; writing–review & editing. **Hanna Kivistö de Souza**: conceptualization; methodology; software; supervision; writing–review & editing. **Mary Grantham O'Brien**: investigation; formal analysis; writing–review & editing. **Angélica Carlet**: investigation; writing–review & editing.

We extend our gratitude to the participants who volunteered for this study. We also wish to thank the anonymous reviewers and Associate Editor Charlie Nagle for their invaluable feedback and guidance.

A one-page Accessible Summary of this article in nontechnical language is freely available in the Supporting Information online and at https://oasis-database.org

Correspondence concerning this article should be addressed to Cesar Teló, Concordia University, Department of Education, 1455 Blvd. De Maisonneuve O., Montréal, QC, Canada H3G 1M8. Email: cesar.telo@concordia.ca

The handling editor for this article was Charlie Nagle.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Author(s). *Language Learning* published by Wiley Periodicals LLC on behalf of Language Learning Research Club, University of Michigan.
DOI: 10.1111/lang.12682

assignment in a pattern resembling the Dunning-Kruger effect despite the controlled nature of the task. Speakers with larger vocabulary size judged their sentence stress assignment as correct more often and showed greater overconfidence and miscalibration. Finally, the assessments of speakers with a background in applied linguistics and/or language teaching were more aligned with listeners' assessments.

Keywords self-assessment; Dunning-Kruger effect; second language pronunciation; sentence stress; individual differences

Introduction

People's ability to assess their own skills has interested researchers for decades, including those in the fields of second language (L2) learning and testing. Similarly to other domains, L2 users seem to be unaware of their language skills and to engage in flawed self-assessment behavior, such that people's self-assessments differs considerably from external measures of performance (most frequently operationalized as listener-based assessments and referred to as other-assessment). Early (Ross, 1998) as well as recent (Li & Zhang, 2021) meta-analytic evidence shows that self- and other-assessment across several L2 domains correlate moderately on average, with self- and other-assessment of speaking being correlated at $r = .44$ (Li & Zhang, 2021). When examining the patterns that characterize the calibration between self- and other-assessment, studies have frequently found that less skilled performers overestimate their abilities (Isbell & Lee, 2022; Saito et al., 2020; Trofimovich et al., 2016)—a phenomenon known as the Dunning-Kruger effect (Kruger & Dunning, 1999).

To date, the most prominent studies on L2 pronunciation self-assessment (i.e., the ability to evaluate one's own L2 pronunciation) have looked at self- versus other-assessment of global pronunciation-related qualities (e.g., Isbell & Lee, 2022; Nagle et al., 2022; Saito et al., 2020; Strachan et al., 2019; Trofimovich et al., 2016; Tsunemoto et al., 2022), especially comprehensibility (ease of understanding) and accentedness (degree of L2 accent), and fewer studies have addressed self-assessment at the segmental (e.g., Dlaska & Krekeler, 2008; Isbell, 2021) and suprasegmental (e.g., Lappin-Fortin & Rye, 2014; O'Brien, 2019) levels. Prior findings have suggested that prosody is particularly difficult for learners to attend to when assessing their own pronunciation (Derwing, 2003; Derwing & Rossiter, 2002; Wrembel, 2015). However, being unable to produce accurate self-assessments of prosodic features is likely to hinder communication given that, if mistakes go unnoticed by the speakers and these mistakes contribute to reduced intelligibility, speakers may continue to produce the prosodic errors and experience similar

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

communication problems in the future. Relatedly, noticing accounts of L2 acquisition posit that L2 development takes place when L2 learners notice and subsequently minimize the gap between their interlanguage and the target form (e.g., Schmidt, 2001). When learners fail to notice this gap because they believe that their productions are targetlike, it is likely that they employ less effort in acquiring the target structure (i.e., minimizing the gap between their interlanguage and the target). The purpose of this study was twofold: (a) To investigate how L2 speakers of English assess their own sentence stress assignment accuracy relative to an external measure of performance (L1 English listeners' assessments); and (b) to examine the extent to which several speaker background characteristics predict self-assessment and the calibration (alignment) between self- and other-assessment.

Background Literature

Self- and Other-Assessment of Second Language Phonological Features

Listener-based assessments are used for several purposes in L2 speech research, including for estimating speakers' intelligibility, comprehensibility, and accentedness, as well as their segmental and suprasegmental accuracy (Saito et al., 2017). Although not immune to biases and idiosyncrasies (Derwing & Munro, 2015), listeners (raters) tend to agree in the speech ratings that they provide (e.g., Munro et al., 2006), allowing other-assessment to be generally considered sufficiently valid and reliable. Prior studies have successfully employed listeners as external assessors of phonological features, including prosody, for a variety of tasks, such as identifying stress assignment, assessing stress assignment accuracy, and evaluating intonational appropriateness and intentionality (e.g., Kivistö de Souza, 2017; O'Brien, 2019; Passarella-Reis et al., 2016).

Assessing the accuracy of one's own L2 pronunciation, on the other hand, may be more challenging than assessing that of others. Emerging evidence demonstrates that L2 users are not fully aware of their pronunciation skills (e.g., Trofimovich et al., 2016), which suggests a lack of metalinguistic awareness, defined as "an individual's ability to focus attention on language as an object in and of itself, to reflect upon language, and to evaluate it" (Thomas, 1988, p. 531). For example, when L2 users identify their own pronunciation problems, segmental issues tend to be the most frequently noticed deviations (Derwing, 2003; Wrembel, 2015), comprising as much as 84% of the detected problems (Derwing & Rossiter, 2002). Conversely, identifying suprasegmental deviations in one's own speech seems to be more challenging: Wrembel's (2015) analyses of stimulated recall protocols show that only

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

1.5% of the comments about learners' own perceived pronunciation problems referred to sentence stress. When comparing self- and other-assessment at the prosodic level, Lappin-Fortin and Rye (2014) reported that only one third of the participants noticed prosodic deviations in their pronunciation but expert raters identified suprasegmental problems in the speech of almost 70% of the participants. Similarly, compared to expert raters' assessments of the location of the stressed syllable in speakers' productions, the speakers in O'Brien's (2019) study provided an accurate assessment of their own lexical stress assignment for only 64% of the test items.

In addition to targeting different phonological features, L2 pronunciation self-assessment studies have employed a variety of instruments to elicit speakers' and listeners' assessments, including Likert scales (e.g., Babaii et al., 2016; Lappin-Fortin & Rye, 2014), think-aloud protocols (e.g., O'Brien, 2019; Wrembel, 2015), interviews (e.g., Derwing, 2003; Isbell, 2021), and yes/no or open-ended questionnaires (e.g., Dlaska & Krekeler, 2008; Meritan & Mroz, 2019). These instruments share the characteristic that they allow participants to carefully reflect upon and monitor their performance, processes that are dependent on L2 users' metalinguistic knowledge and abilities (Roehr-Brackin, 2018). However, less skilled performers may provide inaccurate self-assessments due to a deficit in metacognitive skills (Kruger & Dunning, 1999), meaning that they are less able to reflect on their own skills and produce self-assessments that align with those provided by external assessors. It is presently unknown if prior findings would hold true if self-assessments were elicited without relying so extensively on participants' explicit knowledge, likely required by scales, think-aloud protocols, and interviews. Therefore, employing a judgment task that relies on speakers' implicit knowledge to a greater extent (by means of restricting the time available to judge the stimuli, for example; Plonsky et al., 2020) may generate novel insights into L2 pronunciation self-assessment.

Individual Differences in Self-Assessment

Notwithstanding general patterns of inaccurate self-assessment, researchers have investigated individual differences in self-assessment with the goal of providing insight into the variability in metacognitive abilities among people. In the original examination of the Dunning-Kruger effect (Kruger & Dunning, 1999), variability in performance was posited as a domain-specific phenomenon, implying that a person's performance in intellectual tasks may exhibit great variation across different contexts. This premise initiated a

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

fruitful avenue of research aiming to uncover the specific characteristics that are associated with individuals' ability to recognize their own shortcomings.

In the case of L2 pronunciation, prior studies, albeit limited in number, have revealed that speaker-specific characteristics may be associated with the accuracy of self-assessment, usually defined as the alignment between self- and other-assessment. Language proficiency, experiential variables (e.g., engagement in extracurricular language practice), cognitive variables (e.g., working memory), and people's attitudes towards pronunciation have been considered as relevant background variables in L2 pronunciation self-assessment research. This research has shown, for example, that higher proficiency levels and more extensive L2 use are positively correlated and predict the accuracy of L2 pronunciation (Isbell & Lee, 2022; Li, 2018; Saito et al., 2020). Speakers' demographic characteristics, including speakers' age and L1, may also be related to accuracy in L2 pronunciation self-assessment (Li, 2018; Trofimovich et al., 2016). Lastly, attitudes towards pronunciation have also been shown to predict the calibration between L2 pronunciation self- and other-assessment; specifically, greater calibration has been observed for speakers with greater satisfaction with their own pronunciation and who value pronunciation to a greater extent (Isbell & Lee, 2022).

Sentence Stress Assignment

Sentence stress—also referred to as nuclear stress (Wells, 2006), phrasal stress (O'Brien, 2022), and prominence (Levis, 2018), among other terms—is understood as the last and most phonetically prominent pitch accent within a phrase (Wells, 2006). It is arguably "the most important intonational feature in terms of intelligibility" (Levis, 2018, p. 155), and deciding where it goes is crucial in selecting an intonation pattern (Wells, 2006). Sentence stress is critical because making a syllable the most prominent within a phrase (that is, assigning sentence stress) is a dynamic process related to discourse meaning and information structure, such as signaling the most pragmatically relevant word within an utterance. Several rules govern sentence stress assignment, and placing stress correctly may be especially challenging for L2 speakers, since languages differ in the rules and strategies (phonetic-phonological and/or morphosyntactic) that they employ to highlight specific parts of an utterance (Féry et al., 2010; Krifka, 2008).

Phrases in which all of the information is considered to be new to the listener, and thus no specific part of the phrase is intended to be pragmatically emphasized, are referred to as "all-new," "broad-focus," or "out-of-the-blue." In this type of phrase in English, sentence stress is placed towards the end of

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the phrase (Wells, 2006). Consider, for instance, the following example where the word receiving sentence stress is marked via underlining, and its stressed syllable is marked with ˈ: *She's just started a new re*ˈ*lationship*. In this case, sentence stress is assigned following the default stress rule, which "has no meaning or function: it is simply the result of the operation of phonological rules on surface syntactic structures" (Ladd, 2008, p. 216). However, under two circumstances, sentence stress is placed on an earlier element of the phrase despite it being in broad focus: When the phrase ends in a function word (auxiliary verbs, modal verbs, personal pronouns, prepositions; e.g., *She* ˈ*loves him*), or in old, given, shared information (e.g., *A: Do you object to dogs? B: No, I a*ˈ*dore dogs*; Wells, 2006). Thus, in those cases, sentence stress is assigned to the last content word or to new information—unless special circumstances apply. The target structure selected for this study is broad-focus phrases ending in one or more function words (see below).

Participants in this study were L1 Brazilian Portuguese (BP) speakers. Although English and BP resemble each other in some aspects of sentence stress assignment (e.g., contrastive stress as in *I said to* ˈ*place it down, not* ˈ*throw it down*), certain circumstances require the acquisition of new rules by L1 BP speakers of English. Sentence stress assignment in all-new, broad-focus phrases in BP is much more rigid, with BP placing sentence stress on the last constituent of the phrase by default (Tenani, 2002; Truckenbrodt et al., 2009). In other words, regardless of whether the last element of the phrase is a content word or a function word, it receives sentence stress, as in *Ela ama* ˈ*ele* ("She loves him"). The most common implication of simply transferring this sentence stress assignment rule from BP to English is a shift in focus, in which case a broad-focus phrase is narrowed, denoting contrast. Consider, for instance, the following example from Passarella-Reis (2017, p. 80):

Dê o livro para ˈ*mim* ("Give me the book")
Give the book to ˈme

Following the default sentence stress assignment rule, sentence stress is placed on the last constituent of the BP phrase (*mim*, "me," which happens to be a function word), and no contrast or emphasis is implied. However, if a Brazilian speaker of English applies the BP default rule when producing declarative broad-focus phrases in English, the focus of the phrase is narrowed, and a contrast is intended between *me* and another person. Learning where to assign

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

sentence stress in English declarative broad-focus phrases requires, therefore, the acquisition of a new rule by L1 BP speakers.

The Present Study

In this study, we extended previous work on L2 pronunciation self-assessment by investigating self-assessment of sentence stress assignment accuracy. Speakers' and listeners' assessments were elicited via a timed, speeded judgment task that presented low-pass filtered stimuli (see below), thus encouraging more implicit judgments than tasks adopted in previous studies. We compared BP L2 speakers' self-assessment to the assessment provided by L1 English-speaking listeners. Furthermore, because individual differences may play a role in the extent to which people engage in inaccurate self-assessment (Dunning, 2011) but L2 pronunciation research exploring specific variables is scarce, we conducted an exploratory investigation in order to determine the relative role of several speaker background characteristics in predicting self-assessment and self- and other-assessment calibration. We placed particular emphasis on language- and learning-related variables, given our speakers' diverse experiences. First, we explored whether having a background in applied linguistics and/or language teaching influenced speakers' self-assessment, assuming that such backgrounds might lead to heightened language awareness. Second, we aimed to understand how extended formal language learning might affect self-assessment, considering that individuals with more years of learning likely received more feedback on their pronunciation. English experience was operationalized as the number of years spent studying English in formal educational contexts (2–30 in our dataset). Third, we investigated whether having received previous instruction in English pronunciation and/or phonetics played a role, as such instruction might have provided participants with particular insights into sentence stress or their pronunciation as a whole. Lastly, we examined the role of vocabulary size, which serves as an estimate of proficiency. We chose this final variable because previous research on self-assessment of global pronunciation-related dimensions has shown a relationship with proficiency, and we aimed to extend this to the self-assessment of a discrete phonological feature. The following research questions guided the study:

1. What is the relationship between self-assessment of sentence stress assignment accuracy and the assessment provided by L1 English listeners?
2. To what extent do speaker background characteristics explain differences in self-assessment and in self- and other-assessment calibration?

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

On the basis of the literature describing self-assessment in various domains, including L2 pronunciation, we predicted that speakers, especially those who were least skilled, would produce mostly inaccurate self-assessments and overestimate their performance (Kruger & Dunning, 1999; Dunning, 2011; Trofimovich et al., 2016). Concerning the individual differences analysis, we predicted that speakers with a background in applied linguistics and/or language teaching would produce more accurate self-assessment, as these experiences might foster heightened language awareness (e.g., Andrews, 1999). Conversely, because pronunciation instruction usually places greater emphasis on segments (Isaacs, 2018) and the targeted sentence stress rule rarely features in textbooks (Levis, 2018), we hypothesized that merely having received pronunciation instruction would not necessarily be associated with self-assessment accuracy. Finally, we expected those speakers who engaged with language learning and use contexts more extensively and those with higher proficiency (as estimated by vocabulary size) to tend to produce more accurate self-assessments (Isbell & Lee, 2022; Saito et al., 2020).

Method

Participants

Two groups of participants took part in the study: speakers and listeners. The speakers were 38 L1 BP speakers of English who volunteered to participate in the study through ads on social media and university mailing lists.[1] They were 24 females and 14 males, $M_{age} = 26.79$ years, $SD = 6.46$, minimum = 20, maximum = 45. No minimum level of proficiency in the L2 was required, but potential speakers were informed that the operational language of the instruments was English. Speakers' educational and professional backgrounds were mostly related to applied linguistics and/or language teaching. Nineteen speakers (50%) had either graduated or were majoring in foreign languages and literatures (most of them in English language and literature, but some in English and Portuguese, one in French, and one in Italian), and the remaining had a background in varying fields. Seventeen speakers (44.73%) reported working as language teachers. Their experience with English in formal learning contexts was of 11.92 years on average, $SD = 6.24$, minimum = 2, maximum = 30, and 25 speakers (65.78%) reported having taken a course on English pronunciation and/or phonetics. Speakers' vocabulary size was measured using a receptive vocabulary knowledge test (V_YesNo; Meara & Miralpeix, 2017) as an estimate of their English proficiency. The test requires that the test taker indicates whether they know the meaning of several words that are presented orthographically among nonwords. On average, speakers scored 7,274.24 points out

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

of 10,000, $SD = 972.57$, minimum = 5,556, maximum = 9,136. Scores from 4,500 to 7,500 indicate a learner with "a good level of competence" (Meara & Miralpeix, 2017, p. 118). V_YesNo results have been previously related to L2 speaking proficiency (Uchihara & Clenton, 2020).

The listener group was composed of 29 L1 English-speaking undergraduate students from a university in Canada. They were recruited from a linguistics subject pool and received course credit for their participation. They were 26 women, two men, and one agender, $M_{age} = 21.13$ years, $SD = 5.27$, minimum = 18, maximum = 37. Listeners reported no experience with Portuguese-accented English. All but one reported having normal hearing; because excluding that listener's data resulted in no change in the findings, the entire dataset was used for analysis.[2]

Target Structure

As explained previously, English and BP share some sentence stress assignment rules. Nevertheless, they differ considerably in broad-focus phrases, where BP invariably assigns sentence stress to the last constituent of the phrase and English deaccents the last item if it is a function word or given information.

Three other reasons supported the choice of declarative broad-focus sentences ending in function words. First, this sentence stress pattern is largely regular in English, which allows it to be found in up to 90% of English phrases (Crystal, 1969). Second, this is the most neglected sentence stress use in teaching materials (Levis, 2018), which makes it unlikely that participants received any formal instruction regarding this particular sentence stress assignment rule. Third, L1 BP speakers seem to be more aware of sentence stress in deaccented sentences—especially those deaccented due to ending in function word (e.g., *There's a de'livery for you*)—than in unaccusative sentences (e.g., *New 'evidence emerged*; Kivistö de Souza, 2017), and such sensitivity to the rule was deemed important for completing the assessment task.

As the accuracy of sentence stress assignment can be assessed only within a specific grammatical context, seven question-answer dialogues were created for the study. The questions ensured that the focus was broad, and the answers presented the target sentences (see below). On average, the sentences were 5.28 words long, $SD = 0.95$, minimum = 4, maximum = 7. The vocabulary of the sentences was carefully selected to avoid typically challenging words for L1 BP speakers of English, as complex words might hinder accurate sentence stress placement (Passarella-Reis, 2017). The complete list of stimuli used in the study is available online (see Appendix S1 in the online Supporting Information).

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Materials

Speech Elicitation Task

The target sentences were elicited with a reading task administered online, on Testable (Rezlescu et al., 2020). Speakers were instructed to silently read a contextualizing sentence and a dialogue composed of a question and answer. For example:

> *The intercom rings and you answer it. Your friend asks you*: [context]
> *What's that?* [question]
> *There's a delivery for you*. [answer]

After reading the contextualizing sentence and the dialogue, speakers clicked on a button to start the recording. They were then shown only the target sentence (i.e., the answer) and were asked to read it out loud. The recordings served as the basis for the assessment task.

Self- and Other-Assessment Task

Speakers and listeners assessed sentence stress assignment accuracy via a timed, speeded, forced-choice judgment task. Each trial had the following structure: First, participants read a contextualizing sentence and question as described above (e.g., *The intercom rings and you answer it. Your friend asks you: What's that?*), which remained on the screen for 6,500 ms. Next, the answer to the question (e.g., *There's a delivery for you*.) was presented orthographically and remained on the screen for 2,500 ms. The written answer was then replaced by the image of a loudspeaker that accompanied the recording of the answer that the participants had just read. Finally, participants judged whether the sentence stress was correctly placed by answering the question "Was the stress correctly placed on the sentence?" using the keys "A" (for "yes") and "L" (for "no"). Participants were instructed to click on a key as fast and as accurately as possible. The expectation with this task was that reading the target sentence would trigger the retrieval of the prosodic representation of the phrase from participants' long-term memory, which would then be compared to the low-pass filtered stimulus presented immediately after.

Prior to being used as stimuli for the assessment task, the recorded sentences were low-pass filtered using (Audacity®, 2021). The recordings were submitted to noise reduction, amplification, low-pass filtering at 400 Hz with a roll-off slope of 48 dB per octave, and normalization of peak amplitude to 0.0 dB. Low-pass filtering had the objective of directing participants' attention to the suprasegmental level of the speech signal, as the manipulation removed most of the segmental cues while keeping the suprasegmental information in-

tact. This procedure has been successfully used to allocate listeners' attention to prosody in previous L2 speech research (e.g., Trofimovich & Baker, 2006).

To ensure that sentence stress assignment could be reliably evaluated in the low-pass filtered recordings and to confirm that no target sentence was problematic in this regard, a quality control check was performed. Two L1 English-speaking experts (with doctoral training in phonetics and unfamiliar with the recordings) independently evaluated the recordings by completing a task similar to that completed by the listeners (i.e., judging whether sentence stress was correctly assigned). By-sentence, prevalence-corrected Cohen's kappa coefficients revealed almost perfect agreement between the expert raters overall ($\kappa = .83$). The sentence on which the expert raters disagreed the most yielded a prevalence-corrected Cohen's $\kappa = .63$ (indicative of substantial agreement; Sim & Wright, 2005), and the kappa for the sentence on which they agreed the most was .95 (see Appendix S1 in the online Supporting Information for the kappa coefficient of each target sentence). Therefore, although speech is commonly subject to a great deal of deviation from grammatical (including phonetic-phonological) rules, these results indicate that sentence stress assignment could be reliably perceived and evaluated in the recordings.

Procedures

Speakers first completed the speech elicitation task, followed by the V_YesNo vocabulary size test and a background questionnaire. The questionnaire tapped into speakers' English learning experience, educational background, and occupation, allowing us to derive three background variables: applied linguistics/teaching (a yes/no response regarding whether speakers had an educational and/or professional background in applied linguistics and/or language teaching), experience (number of years spent studying English in formal educational contexts), and pronunciation instruction (whether speakers had received English pronunciation and/or phonetics instruction previously).

Between four and six weeks after the completion of the speech elicitation task, speakers completed the self-assessment task. Speakers and listeners used very similar tasks to assess the accuracy of sentence stress assignment. Speakers first judged the productions of three peers (which are not reported in this manuscript) and then proceeded to assessing their own productions. Prior to beginning the self-assessment block, speakers read a message informing them that they would be subsequently judging their own productions. The speakers' task was administered using Testable (Rezlescu et al., 2020). The listeners' version of the task included all seven trials from the 38 speakers presented in a randomized order and was administered via a jsPsych experiment hosted on

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Cognition (de Leeuw, 2015). Both versions of the task included three practice trials and an explanation of what sentence stress is. The tasks were completed remotely. Both speakers and listeners were instructed to wear headphones and to complete the tasks in a silent environment where they could focus on the tasks.[3]

Data Analysis

Each speaker contributed to the assessment task by uttering seven sentences. Each listener judged on a binary scale whether the sentence stress was correctly assigned in all seven sentences produced by each of the 38 speakers, totaling 266 sentences. Listeners' judgments were first checked for interrater consistency using a two-way, consistency, average-measure intraclass correlation (ICC) computed using the psych package (version 2.3.6; Revelle, 2023) in R (version 4.3.3; R Core Team, 2024). To streamline this process and because the quality control check indicated no outstanding issues with the target sentences, the ICC was constructed to ignore possible variance among the listeners in their judgments of specific target sentences, focusing instead on how consistent the listeners were in their evaluations of each speaker. Therefore, we treated speakers as items evaluated on a 7-point scale, where 7 represented the maximum score that a speaker could obtain from each rater if all productions were judged to be correct. The ICC coefficient was sufficiently high (.93), and a mean accuracy score for each speaker was then calculated and expressed via a percentage, which was used in all subsequent analyses. Therefore, each speaker was attributed two accuracy scores: one for self-assessment and one as assessed by L1 English listeners.

Following prior studies (e.g., Isbell & Lee, 2022), two self- and other-assessment calibration measures were computed: overconfidence and miscalibration. Overconfidence scores were obtained by subtracting the mean other-assessment score from the mean self-assessment score. Positive values indicated that speakers overestimated their sentence stress assignment accuracy, and negative values indicated that speakers underestimated their sentence stress use. Scores around zero indicated alignment between self- and other-assessment. Miscalibration scores referred to the magnitude of the difference between self- and other-assessment and were obtained by calculating the absolute difference between the self- and other-assessment scores. Therefore, miscalibration scores did not qualify the calibration between speakers and listeners in terms of under- or overconfidence since the scores were always a positive number. Scores around zero indicated calibration between self- and other-assessment, and the farther the score was from zero, the greater was

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the mismatch between speakers' and listeners' assessments. In summary, the main study variables were self-assessment, other-assessment (by L1 English listeners), overconfidence, and miscalibration.

The normality of all variables was checked through visual methods (histograms and Q-Q plots) and further confirmed with a hypothesis test (Shapiro-Wilk). Because some variables were not normally distributed, descriptive analyses included median (Mdn) and interquartile range (IQR) values as measures of central tendency and variability. Spearman correlations were used when at least one of the variables at issue was not normally distributed, and Pearson correlations were adopted when both variables were normally distributed. Correlation coefficients were interpreted according to field-specific guidelines, that is, .25 for small, .40 for medium, and .60 for large correlations (Plonsky & Oswald, 2014). In line with prior research (Trofimovich et al., 2016), the calibration scores were also ranked into thirds on the basis of speakers' performance as judged by the listeners, and the bottom and top thirds were compared using two-sample *t* tests.

We fitted multiple linear regression models to investigate the potential of speaker individual differences in explaining self-assessment and self- and other-assessment calibration. Each model included the following variables as predictors: (a) other-assessment (continuous); (b) background in applied linguistics and/or language teaching (binary); (c) L2 experience (number of years spent studying English in formal contexts; continuous); (d) pronunciation instruction (whether speakers had received English pronunciation and/or phonetics instruction previously; binary); and (e) vocabulary size (continuous). All variables were standardized prior to entering the models. Binary variables were centered, and continuous variables were centered and divided by two standard deviations (Gelman & Hill, 2007; Sonderegger, 2023). Collinearity among variables was checked via variance inflation factors and was not found to be a concern in any model (< 1.5). Histograms, Q-Q plots, residual plots, and Shapiro-Wilk tests were used to assess the normality of model residuals. Overfitting (the possibility that the model does not generalize well to unobserved data) was assessed by comparing R^2 and adjusted R^2 values (Winter, 2020). All models appeared to overfit to a certain extent. Although it is likely that the relatively large number of predictors accounts at least partially for the overfitting, our a priori decision was to conduct exploratory regression analyses given the gap in the literature concerning individual differences that might be related to self-assessment and to self- and other-assessment calibration of L2 discrete phonological features. Furthermore, adopting methods to control for overfitting would introduce bias (Sonderegger, 2023), so predictors were

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1 Descriptive statistics for self- and other-assessment of sentence stress assignment and calibration measures

Variable	*M*	*SD*	95% CI	Mdn	Min	Max	IQR
Self-assessment	75.06	16.88	[69.70, 80.43]	71.43	28.57	100	26.19
Other-assessment	57.00	15.55	[52.05, 61.94]	59.61	23.15	78.32	22.80
Overconfidence	18.06	16.82	[12.72, 23.41]	21.18	−16.75	51.23	19.46
Miscalibration	21.20	12.50	[17.23, 25.18]	21.18	0.99	51.23	14.90

Note. $N = 38$.

not dropped and robust regressions were not conducted. Finally, no clear outliers were found using the Bonferroni outlier test (car package; version 3.0.13; Fox & Weisberg, 2019). The data are publicly available through an Open Science Framework study profile (http://doi.org/10.17605/OSF.IO/8Z739).

Results

Table 1 summarizes the perceived accuracy in sentence stress assignment, in percentage, as assessed by speakers and listeners, as well as the calibration measures (overconfidence and miscalibration). When comparing self- and other-assessment, the mean accuracy values were different, $M_{\text{self}} = 75.06$ versus $M_{\text{other}} = 57.00$, suggesting that, as a group, speakers were overconfident in their assessments of sentence stress assignment accuracy. Nevertheless, the calibration measures show that, as individuals, participants displayed a considerable amount of variation in calibration between self- and other-assessment.

A Spearman correlation between self- and other-assessment showed that speakers' judgments were positively and moderately correlated with listeners' assessments, $\rho = .45$, $p = .005$, as illustrated in Figure 1.

Next, listeners' ratings were correlated with the calibration measures to investigate the relationship between actual performance and overconfidence/miscalibration. The correlation between listeners' ratings and overconfidence scores was negative and moderate, $\rho = -.46$, $p = .004$, indicating that less accurate sentence stress assignment was associated with greater overconfidence. A very similar result was obtained when listeners' ratings were correlated with miscalibration scores, which are absolute differences between self- and other-assessment, $\rho = -.45$, $p = .005$. In other words, the speakers whose sentence stress assignment was judged as being the least accurate by L1 English listeners were also the ones who were the most overconfident and miscalibrated in their self-assessment of sentence stress assignment accuracy.

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

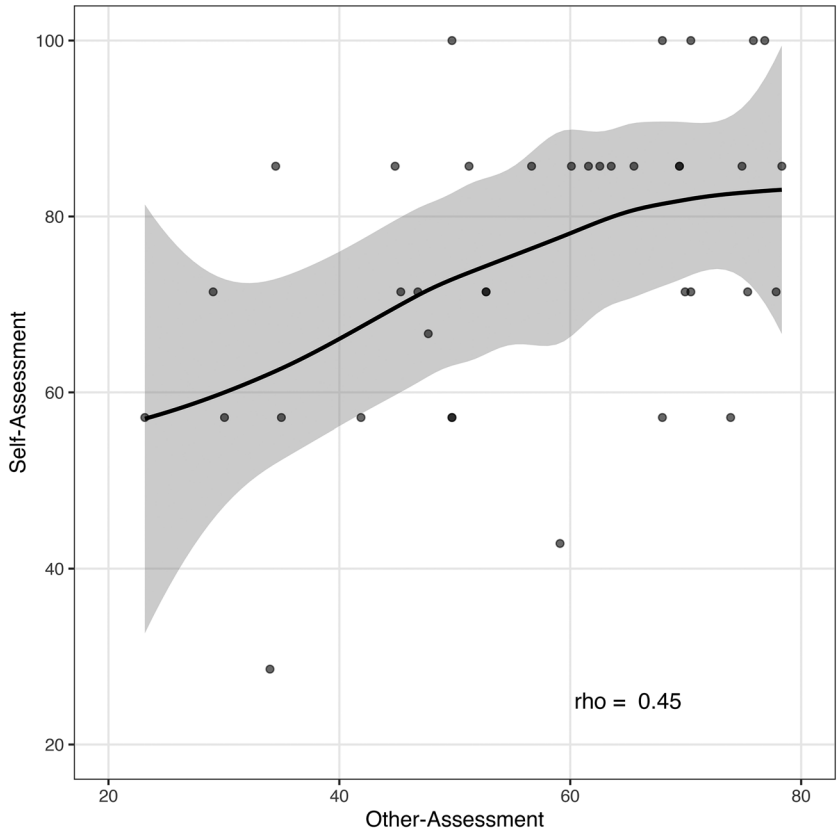

Figure 1 Scatterplot depicting the association between speaker- and listener-rated sentence stress assignment accuracy (in percentage). Trends are indicated by a LOESS line and Standard Error (*SE*)

The scatterplots in Figure 2 show the relationship between other-assessment and the calibration measures. The locally estimated scatterplot smoothing (LOESS) lines depict the best fit to the data.

To obtain a more fine-grained perspective on the relationships described above, the overconfidence and miscalibration scores were split into thirds according to listener-based assessment. The overconfidence and miscalibration scores for the bottom and top thirds of the speakers—as rated by the listeners—were then compared using *t* tests (two-sample, unequal variances). For overconfidence, the bottom third, $M = 27.03$, $n = 11$, was significantly more overconfident than the top third, $M = 8.58$, $n = 12$, $t(20.32) = 3.01$,

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

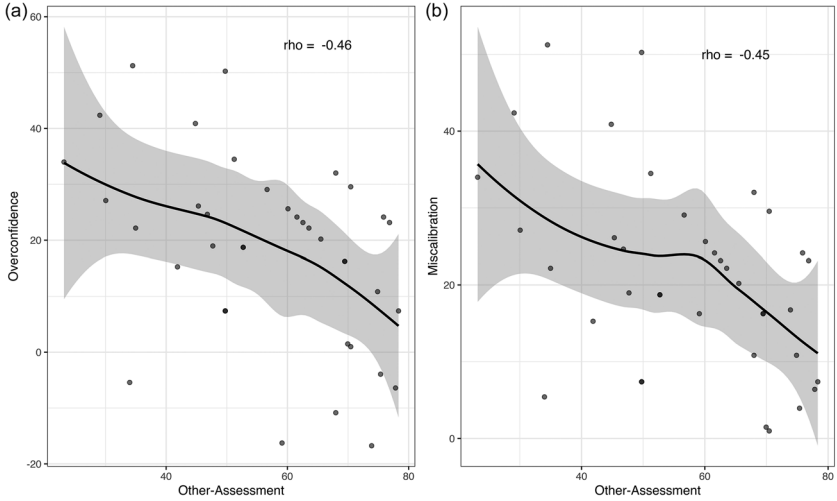

Figure 2 Scatterplots illustrating the relationship between listeners' ratings and calibration measures: overconfidence (panel A) and miscalibration (panel B)

$p = .007$, $d = 1.26$, 95% CI [0.34, 2.15]. For miscalibration (the absolute difference between self- and other-assessment), the bottom third, $M = 28.01$, $n = 11$, was again significantly more miscalibrated (and overconfident) than the top third, $M = 13.09$, $n = 12$, $t(17.92) = 3.09$, $p = .006$, $d = 1.30$, 95% CI [0.36, 2.21], meaning that the group of speakers whose sentence stress assignment was the most incorrect (or less often judged as correct according to L1 listeners) was reliably more overconfident and miscalibrated than the group of speakers who was judged to be the best at sentence stress assignment. All mean differences between groups are considered large according to L2 research guidelines (Plonsky & Oswald, 2014).

Finally, to further qualify how the different groups behaved, speakers were grouped into performance quartiles on the basis of how they were assessed by the listeners. Figure 3 shows self- and other-assessment (in dashed and solid lines, respectively), plotted along listener-rated performance quartiles. As Figure 3 shows, speakers whose sentence stress assignment accuracy was the lowest as per listeners' judgment (the bottom 25% of the sample) were those who overestimated their pronunciation the most. Conversely, speakers whose sentence stress placement was judged as correct in most of the sentences (the top 25% of the sample) tended to underestimate their pronunciation relative to the assessment of others. Speakers around the median judged their own

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

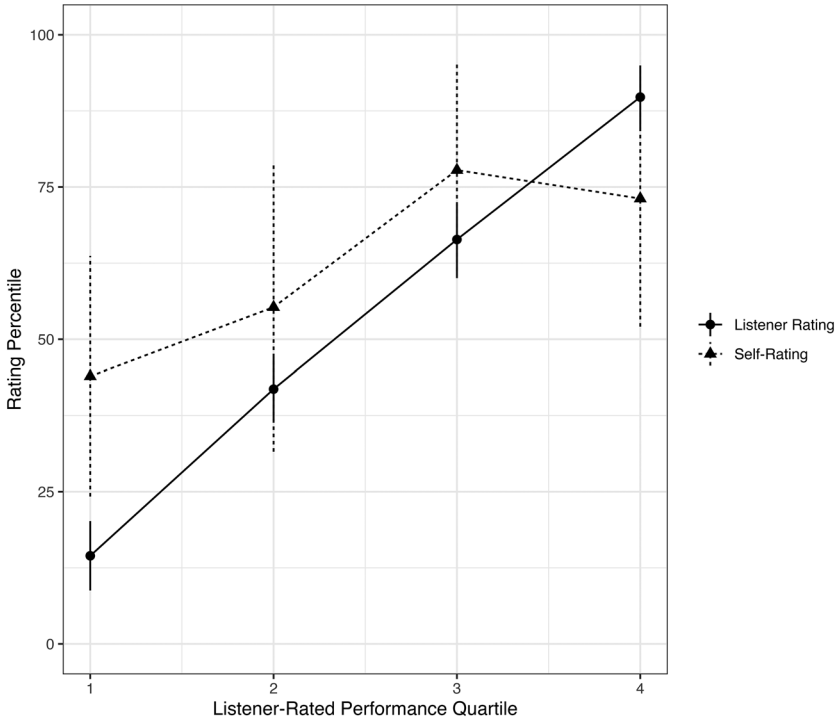

Figure 3 Mean self- and other-assessment percentile ranks by listener-rated performance quartile. Error bars enclose 95% CIs

sentence stress assignment accuracy more similarly to the listeners; nonetheless, self-assessment was still inflated.

To address the second research question, speakers' background variables were loaded into three multiple linear regression models for self-assessment scores, overconfidence, and miscalibration. Following prior research (Isbell & Lee, 2022), the predictors included in each model were associated both with the listener (L1 English listeners' assessment) and with the speaker (background in applied linguistics and/or language teaching, L2 experience, pronunciation instruction, vocabulary size). Table 2 summarizes the regression models.

All models revealed that listeners' assessment significantly predicted the outcome variable. Self-assessment was additionally (and more substantially) predicted by speakers' vocabulary size, an estimate of their English proficiency, such that those who obtained higher scores in the vocabulary size test also perceived their sentence stress assignment as accurate on more occasions.

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 2 Summary of the regression models for self-assessment, overconfidence, and miscalibration

	Self-Assessment		Overconfidence		Miscalibration	
Predictor	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Intercept	0.00	[−0.14, 0.14]	0.00	[−0.14, 0.14]	0.00	[−0.15, 0.15]
Other-Assessment	**0.38**	[0.06, 0.71]	**−0.55**	[−0.88, −0.22]	**−0.55**	[−0.85, −0.24]
AL/Teaching	−0.23	[−0.57, 0.10]	−0.24	[−0.57, 0.10]	**−0.32**	[−0.64, −0.00]
Experience	−0.19	[−0.51, 0.13]	−0.19	[−0.51, 0.13]	−0.18	[−0.48, 0.13]
Pron. Instruction	0.22	[−0.08, 0.52]	0.22	[−0.08, 0.53]	0.20	[−0.08, 0.49]
Vocabulary Size	**0.39**	[0.07, 0.70]	**0.39**	[0.07, 0.71]	**0.43**	[0.13, 0.72]
	$R^2 = .413$		$R^2 = .408$		$R^2 = .475$	
	$RSE = 0.418$ ($df = 32$)		$RSE = 0.419$ ($df = 32$)		$RSE = 0.395$ ($df = 32$)	

Note. $N = 38$. AL/Teaching = background in applied linguistics and/or language teaching; Pron. Instruction = previous pronunciation instruction; *RSE* = relative standard error. Values in boldface indicate statistically significant predictors at the $< .05$ level.

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

This model explained about 32% of the variance in self-assessment, adjusted $R^2 = .321$. Both calibration measures (overconfidence, miscalibration) were also associated with speakers' vocabulary size, meaning that a larger vocabulary size (an indication of higher L2 proficiency) was related to more overconfidence and miscalibration. Furthermore, having a background in applied linguistics and/or language teaching reliably predicted miscalibration scores, where speakers with this background produced self-assessments more similar to the assessment provided by the listeners. Together, the variables predicted 32%, adjusted $R^2 = .316$, and 39%, adjusted $R^2 = .393$, of the variance in overconfidence and miscalibration scores, respectively. Some violation of residual normality was present in the self-assessment, $p = .057$, and overconfidence, $p = .057$, models, but the miscalibration residuals resembled normal distribution, $p = .398$.

Discussion

Self- and Other-Assessment

In the present study, we extended prior research on L2 pronunciation self-assessment by examining how L2 English speakers assess their own sentence stress assignment accuracy. The first research question targeted the association between self-assessment of sentence stress assignment accuracy and the assessment provided by L1 English-speaking listeners. As a group, speakers produced self-assessments that were not perfectly aligned with how listeners assessed them; more specifically, speakers' self-assessments were somewhat inflated when compared to the external measures of performance. This means that, overall, speakers thought that they had assigned sentence stress correctly more often (to more sentences) than did the listeners. Correlational analyses revealed that speakers' and listeners' assessments were only moderately correlated, $\rho = .45$. Broadly speaking, our findings aligned with prior work on L2 pronunciation self-assessment (e.g., Saito et al., 2020; Trofimovich et al., 2016) inasmuch as there was a mismatch between self- and other-assessment. However, the magnitude of the relationship between the target variables in our study is larger than those found by Saito et al. (2020) and Trofimovich et al. (2016). In fact, our results resemble those obtained by Isbell and Lee (2022), Li (2018), and Lappin-Fortin and Rye (2014), who reported medium-strength correlations between the self- and other-assessment of both global dimensions of L2 speech and discrete phonological features.

When analyzed more closely, clear patterns emerged from speakers' self-assessments: (a) poor performers (the bottom 25% as per listeners' assessment) overestimated their performance; (b) top performers (the top 25% as

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

per listeners' assessment) underestimated their performance; and (c) "average" performers (middle 50% as per listeners' assessment) assessed themselves more similarly to how listeners assessed them, although self-assessments were still somewhat overconfident (see Figure 3). These patterns resonate with the Dunning-Kruger effect and prior research on L2 pronunciation self-assessment showing the tendency of low-skill speakers to overestimate their pronunciation at the same time that high-skill speakers often underestimate their abilities (e.g., Isbell & Lee, 2022; Saito et al., 2020; Trofimovich et al., 2016). It is noteworthy that only the speakers in the top performance quartile in our study were underconfident relative to listeners' judgements. In previous L2 pronunciation self-assessment studies, speakers in the third (Trofimovich et al., 2016) and even the second (Li, 2018) performance quartiles were found to underestimate their performance. Although this discrepancy could stem from the nature of the data, which were obtained from the judgment of seven sentences per speaker, it could also stem from the fact that people tend to be more confident when self-assessments are elicited more quickly rather than more slowly (e.g., Dunning & Stern, 1994), indicating an effect of the instrument adopted. These judgments are arguably more ecologically valid since they were elicited more quickly, allowing less reflection and, consequently, more implicit accuracy judgments to appear.

Following prior research on self-assessment, we computed two calibration measures to further qualify the association between self- and other-assessment: overconfidence (self-assessment minus other-assessment) and miscalibration (the absolute difference between self- and other-assessment). Both calibration measures yielded similar results: Speakers whose sentence stress assignment was judged as inaccurate on more occasions (sentences) by L1 English listeners were also the ones who were more overconfident, $\rho = -.46$, and miscalibrated, $\rho = -.45$. These results provide additional evidence of the Dunning-Kruger effect and align well with those of other studies on L2 pronunciation self-assessment that obtained negative, medium-strength associations between actual performance (other-assessment) and calibration measures (e.g., Li, 2018).

In summary, the results discussed so far indicate that, as a group, speakers perceived their own prominence assignment accuracy differently from external assessors and tended to overestimate their own pronunciation. Notably, speakers in this study were unable to perfectly assess their performance despite being asked to attend to their sentence stress assignment only and having their attention restrained to the suprasegmental level by means of judging low-pass filtered stimuli. Furthermore, although flawed self-assessment is

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

argued to stem from deficits in metacognitive skills (Kruger & Dunning, 1999)—which involves careful reflection upon performance (Roehr-Brackin, 2018)—the self-assessment task employed was fast-paced, arguably restricting speakers' access to their declarative knowledge about L2 pronunciation (Plonsky et al., 2020). Therefore, adopting a timed self-assessment task that taps into speakers' implicit knowledge seems to have had no effect in terms of helping speakers produce more calibrated self-assessments. Nevertheless, the speakers were aware that they were rating their own productions, which may have triggered other psychological processes that could explain at least partially the discrepancy between self- and other-assessment (e.g., speakers may have been aware that their performance was being assessed, which potentially led them to judge their productions more positively overall).

Individual Differences in Self-Assessment

The second research question investigated the extent to which self-assessment and self- and other-assessment calibration were predicted by speakers' background characteristics. To our knowledge, this is the first study that investigates the relative contribution of individual differences in self-assessment of a discrete phonological feature. Three exploratory multiple regression analyses were conducted including other-assessment, background in applied linguistics and/or language teaching, L2 experience, previous pronunciation instruction, and vocabulary size as predictors. Results suggest that two background variables may be particularly relevant in predicting self-assessment and self- and other-assessment calibration of sentence stress assignment accuracy: speakers' vocabulary size (an estimate of overall proficiency) and having a background in applied linguistics and/or language teaching. L1 English listeners' assessments (other-assessment) emerged as a significant predictor in all three models. This was expected at least for the calibration models since the independent variables were partially derived from the L1 other-assessment variable.

Self-assessment was most substantially predicted by vocabulary size, indicating that the larger a speaker's vocabulary size was, the higher was their mean self-assessment rating. This finding seems to suggest a partial independence between overall proficiency (as estimated by a vocabulary size test) and self-assessed pronunciation skills, as most speakers (about 84%) produced overconfident self-assessments (see Figure 3). Alternatively, it could indicate that speakers become more confident as language proficiency increases. Regardless of the interpretation, results resonate with prior research that showed a similar association between proficiency measures and

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

self-assessments of comprehensibility and accentedness (Isbell & Lee, 2022; Suzukida, 2024). Confirming the results of the correlational analysis between self- and other-assessment, listeners' assessments were positively related to speakers' self-assessments, indicating that, to a certain extent, listeners and speakers agreed in the assessments that they provided, at least distributionally.

As expected, the L1 overconfidence and L1 miscalibration models patterned similarly, with a strong association between the two variables, $r = .88$, $p < .001$. Both calibration models were most substantially predicted by L1 listeners' assessment (L1 other-assessment), but, inversely from the self-assessment model, greater overconfidence and greater miscalibration were associated with lower other-assessed accuracy. Put differently, speakers whose listener-assessed accuracy was low were reliably more overconfident and miscalibrated, providing additional evidence for the Dunning-Kruger effect. Vocabulary size also emerged as a significant predictor in both calibration models, where a larger vocabulary size was associated with greater overconfidence and miscalibration. This finding is partially aligned with Isbell and Lee (2022) and Suzukida (2024), who also reported that speakers with higher proficiency (as estimated through an elicited imitation test and through a vocabulary size test, respectively) were more overconfident in their pronunciation self-assessments. Li (2018), on the other hand, observed an opposite interaction between L2 proficiency (as estimated by Test of English as a Foreign Language speaking and listening scores) and overconfidence, although shown by correlational analyses.

Background in applied linguistics and/or language teaching significantly predicted miscalibration scores: Speakers with this background (that is, those majoring in applied linguistics, those who had graduated from an applied linguistics program, and/or those who worked as language teachers) were more calibrated. It is likely that having formally studied and/or taught second language(s) was associated with higher levels of linguistic and phonological (self-)awareness, which resulted in self-assessments that were more similar to the assessment of L1 English listeners (Andrews, 1999; Kennedy & Trofimovich, 2010; Moyer, 2014; see also Griffin et al., 2009). Relatedly, it is possible that the speakers who had applied linguistics and/or English teaching experience were better at assessing their own skills because they had been exposed to a greater extent to people whose sentence stress assignment is accurate—although closely attending to other people's speech is also related to (meta)cognitive skills (e.g., Moyer, 2014). Regardless of the explanation, this finding is of key importance, as it may render evidence for noticing accounts of L2 acquisition, in which L2 development is argued to take place

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

when L2 speakers notice and subsequently minimize the gap between their interlanguage and the target form (e.g., Schmidt, 2001).

Two variables did not emerge as significant predictors in any model: L2 experience and previous English pronunciation instruction. Although we predicted that speakers would present flawed self-assessment behavior, it is noteworthy that, on average, speakers had almost 12 years of experience studying English in formal learning context, which certainly had provided them with numerous opportunities to compare their output with the input received and to receive feedback on their pronunciation. Yet, despite a quantitively large amount of experience learning English, speakers learned English in a foreign language setting, where input conditions are often less-than-ideal for a variety of reasons: The L2 is restricted mostly to instructional settings (Muñoz, 2008); pronunciation is rarely integrated into the language curriculum (Pennington, 2021); and focus on (linguistic) form may be limited due to the teaching approaches largely adopted (Darcy et al., 2021). Furthermore, the self-assessments of speakers who had received pronunciation instruction were not more aligned with listeners in the assessments that they produced. This might reflect the fact that, despite a general absence of pronunciation instruction from the L2 classroom, when pronunciation is addressed, it tends to be primarily segment-based (Isaacs, 2018). Alternatively, it is possible that the speakers who received previous pronunciation instruction possess increased awareness and explicit knowledge of the L2 phonological system, but, considering that the sentence stress rule investigated in this study is mostly absent from teaching materials (Levis, 2018), speakers had likely never been taught about this particular rule.

Limitations, Future Work, and Implications

Although generally aligned with prior research on L2 pronunciation self-assessment, the present study investigated self-assessment of a single sentence stress assignment rule. As such, the findings reported may not extend to other suprasegmental features, nor to other sentence stress assignment rules. Furthermore, speakers' sentence stress assignment accuracy was assessed in seven sentences, meaning that each accurate sentence (from a sentence stress assignment perspective) contributed to about 14 points of the maximum score of 100 that each participant could obtain. Considering that speakers' and listeners' assessments differed by around 18 points on average, speakers and listeners disagreed in relation to the accuracy of sentence stress assignment on no more than one sentence, on average. Although we expect listeners' judgments of the phonological structure of their L1 to be fully developed, it is

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

also possible that both speakers and listeners provided inaccurate judgments. Moreover, the assessment task adopted presented low-pass filtered stimuli, which calls for caution in the generalization of the findings, considering the increased level of difficulty posed by such an audio manipulation. Another possible limitation is the compressed audio format used by the online platform to save the recordings and the laptop/headphones microphones used by the speakers to complete the speech elicitation task. Nonetheless, online speech data collection took place on the assumption that F0, the main phonetic correlate of prominence (van Heuven & Turk, 2021), seems to be resistant to audio compression and microphone effects (Cavalcanti et al., 2023). Finally, the background characteristics examined here concerned only language- and learning-related variables. Since inaccurate self-assessment is a psychological phenomenon, other psychologically related variables, such as perfectionism and anxiety, could have been powerful in explaining self-assessment and self- and other-assessment calibration (see, e.g., Dunning, 2011). In the future, researchers may also consider investigating the relative contribution of demographic variables (e.g., gender) to self-assessment.

Inaccurate pronunciation self-assessment is likely to have consequences for L2 learners, who may be particularly prone to overlooking the gap between their own pronunciation and the target and, consequently, delay the acquisition of certain phonological features. Given the importance of sentence stress for intelligibility (Levis, 2018) and the still underinvestigated issue of L2 pronunciation self-assessment, researchers may wish to (a) investigate other sentence stress assignment rules as well as other discrete phonological features that impact communication; (b) expand the testing methodologies used to elicit self- and other-assessment; (c) examine other possible ways to minimize the discrepancy between speakers' and listeners' assessment; and (d) investigate a larger and more diverse number of both speaker and listener variables that may help explain L2 pronunciation self-assessment. For teachers, although consensus on the extent of the Dunning-Kruger effect remains elusive, evidence that people overestimate their skills is robust (e.g., León et al., 2023; Li & Zhang, 2021). Implementing benchmarking techniques (Tsunemoto et al., 2022) and providing feedback (León et al., 2023) may aid learners in producing more accurate self-assessments. Additionally, fostering metalinguistic awareness through consciousness-raising activities and content instruction can help reduce overestimation and promote the development of more accurate and intelligible L2 pronunciation (León et al., 2023; Schmidt, 2001). Finally, supporting learners' engagement in meaningful extracurricular language use opportunities may also enhance the alignment be-

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

tween how a speaker perceives their own pronunciation and how it is perceived by others (Saito et al., 2020).

Conclusion

In this study, we investigated the association between self-assessment of sentence stress assignment accuracy by L2 English speakers (L1 BP) as it relates to the assessment provided by L1 English-speaking listeners. Using a fast-paced judgment task that presented low-pass filtered stimuli—an assessment task that arguably allows for more implicit judgments than those adopted in previous L2 pronunciation self-assessment studies—we showed that speakers generally produced self-assessments that resonate with the Dunning-Kruger effect, such that less skilled speakers (as per listeners' judgment) tended to overestimate their sentence stress assignment accuracy. On the other hand, speakers in the top performance quartile underestimated their pronunciation relative to listeners' assessment. Furthermore, results of multiple regression analyses revealed that the alignment between speakers' and listeners' assessments was predicted by speakers' vocabulary size and by whether they had a background in applied linguistics and/or language teaching. In summary, speakers with a larger vocabulary size were more overconfident and miscalibrated, and speakers with the aforementioned educational/professional background were more calibrated in their self-assessments.

Final revised version accepted 20 August 2024

Open Research Badges

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data are available at http://doi.org/10.17605/OSF.IO/8Z739.

Notes

1 Forty-one speakers participated in the study originally. Data from three speakers had to be excluded from the dataset; in the case of one speaker, this was due to insufficient proficiency to complete the tasks in English; in another case, the exclusion was due to poor sound quality, and yet another was due to an error in one of the tasks completed.

2 Initially, more students participated in the study as listeners. However, data from listeners who failed to complete at least 95% of the assessment task were excluded from the dataset.

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

3 A portion of the data reported in this manuscript was collected and published as part of the first author's undergraduate thesis.

References

Andrews, S. (1999). Why do L2 teachers need to 'know about language'? Teacher metalinguistic awareness and input for learning. *Language and Education*, *13*(3), 161–177. https://doi.org/10.1080/09500789908666766

Audacity Team. (2021). *Audacity*®: *Free audio editor and recorder* (Version 2.4.2). [Computer software]. https://audacityteam.org/.

Babaii, E., Taghaddomi, S., & Pashmforoosh, R. (2016). Speaking self-assessment: Mismatches between learners' and teachers' criteria. *Language Testing*, *33*(3), 411–437. https://doi.org/10.1177/0265532215590847

Cavalcanti, J. C., Englert, M., Oliveira Jr, M., & Constantini, A. C. (2023). Microphone and audio compression effects on acoustic voice analysis: A pilot study. *Journal of Voice*, *37*(2), 162–172. https://doi.org/10.1016/j.jvoice.2020.12.005

Crystal, D. (1969). *Prosodic systems and intonation in English*. Cambridge Studies in Linguistics.

Darcy, I., Rocca, B., & Hancock, Z. (2021). A window into the classroom: How teachers integrate pronunciation instruction. *RELC Journal*, *52*(1), 110–127. https://doi.org/10.1177/0033688220964269

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*, 1–12. https://doi.org/10.3758/s13428-014-0458-y

Derwing, T. M. (2003). What do ESL students say about their accents? *Canadian Modern Language Review*, *59*(4), 547–567. https://doi.org/10.3138/cmlr.59.4.547

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins. https://doi.org/10.1075/lllt.42

Derwing, T. M., & Rossiter, M. J. (2002). ESL learners' perceptions of their pronunciation needs and strategies. *System*, *30*(2), 155–166. https://doi.org/10.1016/S0346-251X(02)00012-X

Dlaska, A., & Krekeler, C. (2008). Self-assessment of pronunciation. *System*, *36*(4), 506–516. https://doi.org/10.1016/j.system.2008.03.003

Dunning, D. (2011). The Dunning-Kruger effect: On being ignorant of one's own ignorance. In J. M. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology* (Vol. 44, pp. 247–296). Academic Press. https://doi.org/10.1016/B978-0-12-385522-0.00005-6

Dunning, D., & Stern, L. B. (1994). Distinguishing accurate from inaccurate eyewitness identifications via inquiries about decision processes. *Journal of*

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Personality and Social Psychology, *67*(5), 818–835. https://doi.org/10.1037/0022-3514.67.5.818

Féry, C., Skopeteas, S., & Hörnig, R. (2010). Cross-linguistic comparison of prosody, syntax and information structure in a production experiment on localising expressions. *Transactions of the Philological Society*, *108*(3), 329–351. https://doi.org/10.1111/j.1467-968X.2010.01240.x

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression*. Sage Publishing.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Griffin, T. D., Jee, B. D., & Wiley, J. (2009). The effects of domain knowledge on metacomprehension accuracy. *Memory & Cognition*, *37*, 1001–1013. https://doi.org/10.3758/MC.37.7.1001

Isaacs, T. (2018). Shifting sands in second language pronunciation teaching and assessment research and practice. *Language Assessment Quarterly*, *15*(3), 273–293. https://doi.org/10.1080/15434303.2018.1472264

Isbell, D. R. (2021). Can the test support student learning? Validating the use of a second language pronunciation diagnostic. *Language Assessment Quarterly*, *18*(4), 331–356. https://doi.org/10.1080/15434303.2021.1874382

Isbell, D. R., & Lee, J. (2022). Self-assessment of comprehensibility and accentedness in second language Korean. *Language Learning*, *72*(3), 806–852. https://doi.org/10.1111/lang.12497

Kennedy, S., & Trofimovich, P. (2010). Language awareness and second language pronunciation: A classroom study. *Language Awareness*, *19*(3), 171–185. https://doi.org/10.1080/09658416.2010.486439

Kivistö de Souza, H. (2017). Examining L1 Brazilian Portuguese speakers' sensitivity to English nuclear stress assignment. *Revista de Estudos da Linguagem*, *25*(2), 483–514. http://doi.org/10.17851/2237-2083.25.2.483-514

Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica*, *55*(3/4), 243–276. https://doi.org/10.1556/aling.55.2008.3-4.2

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Ladd, D. R. (2008). *Intonational phonology*. Cambridge University Press. https://doi.org/10.1017/CBO9780511808814

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*(2), 300–320. https://doi.org/10.1111/flan.12083

León, S. P., Panadero, E., & García-Martínez, I. (2023). How accurate are our students? A meta-analytic systematic review on self-assessment scoring accuracy.

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Educational Psychology Review, *35*(4), Article 106. https://doi.org/10.1007/s10648-023-09819-0

Levis, J. M. (2018). *Intelligibility, oral communication, and the teaching of pronunciation*. Cambridge University Press. https://doi.org/10.1017/9781108241564

Li, M. [Mushi]. (2018). *Know thyself? Self- vs. other-assessment of second language pronunciation* (Publication No. 10689352) [Doctoral dissertation, Boston University]. ProQuest Dissertations Publishing. https://hdl.handle.net/2144/27484

Li, M. [Minzi], & Zhang, X. (2021). A meta-analysis of self-assessment and language performance in language testing and assessment. *Language Testing*, *38*(2), 189–218. https://doi.org/10.1177/0265532220932481

Meara, P., & Miralpeix, I. (2017). *Tools for researching vocabulary*. Multilingual Matters. https://doi.org/10.21832/9781783096473

Meritan, C., & Mroz, A. (2019). Impact of self-reflection and awareness-raising on novice French learners' pronunciation. *Foreign Language Annals*, *52*(4), 798–821. https://doi.org/10.1111/flan.12429

Moyer, A. (2014). Exceptional outcomes in L2 phonology: The critical factors of learner engagement and self-regulation. *Applied Linguistics*, *35*(4), 418–440. https://doi.org/10.1093/applin/amu012

Muñoz, C. (2008). Symmetries and asymmetries of age effects in naturalistic and instructed L2 learning. *Applied Linguistics*, *29*(4), 578–596. https://doi.org/10.1093/applin/amm056

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition*, *28*(1), 111–131. https://doi.org/10.1017/S0272263106060049

Nagle, C., Trofimovich, P., Tekin, O., & McDonough, K. (2022). Framing second language comprehensibility: Do interlocutors' ratings predict their perceived communicative experience? *Applied Psycholinguistics*, *44*(1), 131–156. https://doi.org/10.1017/S0142716423000073

O'Brien, M. G. (2019). Attending to second language lexical stress: Exploring the roles of metalinguistic awareness and self-assessment. *Language Awareness*, *28*(4), 310–328. https://doi.org/10.1080/09658416.2019.1625912

O'Brien, M. G. (2022). Making the teaching of suprasegmentals accessible. In J. M. Levis, T. M. Derwing, & S. Sonsaat-Hegelheimer (Eds.), *Second language pronunciation: Bridging the gap between research and teaching* (pp. 85–106). Wiley-Blackwell.

Passarella-Reis, L. (2017). *What do you mean? Nuclear stress in English as an international language: Uses and interpretations* [Doctoral dissertation, Universidade Federal de Santa Catarina]. Repositório Institucional. https://repositorio.ufsc.br/xmlui/handle/123456789/182807

Passarella-Reis, L., Gonçalves, A. R., & Silveira, R. (2016). Perception of intonational patterns and speaker's intentionality in English yes-no questions produced by

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Brazilians. *Revista de Estudos da Linguagem*, *24*(1), 65–97. http://doi.org/10.17851/2237-2083.24.1.65-97

Pennington, M. C. (2021). Teaching pronunciation: The state of the art 2021. *RELC Journal*, *52*(1), 3–21. https://doi.org/10.1177/00336882211002283

Plonsky, L., Marsden, E., Crowther, D., Gass, S. M., & Spinner, P. (2020). A methodological synthesis and meta-analysis of judgment tasks in second language research. *Second Language Research*, *36*(4), 583–621. https://doi.org/10.1177/0267658319828413

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

R Core Team. (2024). *R: A language and environment for statistical computing* (Version 4.3.3) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Revelle, W. (2023). *psych: Procedures for psychological, psychometric, and personality research*. (R package; Version 2.4.6) [Computer software]. https://CRAN.R-project.org/package=psych.

Rezlescu, C., Danaila, I., Miron, A., & Amariei, C. (2020). More time for science: Using Testable to create and share behavioral experiments faster, recruit better participants, and engage students in hands-on research. In B. L. Parkin (Ed.), *Progress in brain research: Real-world applications in cognitive neuroscience* (Vol. 253, pp. 243–262). https://doi.org/10.1016/bs.pbr.2020.06.005

Roehr-Brackin, K. (2018). *Metalinguistic awareness and second language acquisition*. Routledge.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing*, *15*(1), 1–20. https://doi.org/10.1177/026553229801500101

Saito, K., Trofimovich, P., Abe, M., & In'nami, Y. (2020). Dunning-Kruger effect in second language speech learning: How does self-perception align with other perception over time? *Learning and Individual Differences*, *79*, Article 101849. https://doi.org/10.1016/j.lindif.2020.101849

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, *38*(4), 439–462. https://doi.org/10.1093/applin/amv047

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 1–32). Cambridge University Press.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, *85*(3), 257–268. https://doi.org/10.1093/ptj/85.3.257

Sonderegger, M. (2023). *Regression modeling for linguistic data*. MIT Press.

Strachan, L., Kennedy, S., & Trofimovich, P. (2019). Second language speakers' awareness of their own comprehensibility: Examining task repetition and

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

self-assessment. *Journal of Second Language Pronunciation*, *5*(3), 347–373. https://doi.org/10.1075/jslp.18008.str

Suzukida, Y. (2024). Delving into L2 learners' perspective: Exploring the role of individual differences in self-evaluation of L2 speech learning. *Languages*, *9*(3), Article 109. https://doi.org/10.3390/languages9030109

Tenani, L. E. (2002). *Domínios prosódicos no português do Brasil: implicações para a prosódia e para a aplicação de processos fonológicos* [Doctoral dissertation, Universidade Estadual de Campinas]. Repositório da Produção Científica e Intelectual da Unicamp. https://hdl.handle.net/20.500.12733/1592665

Thomas, J. (1988). The role played by metalinguistic awareness in second and third language learning. *Journal of Multilingual and Multicultural Development*, *9*(3), 235–246. https://doi.org/10.1080/01434632.1988.9994334

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., Kennedy, S., Saito, K., & Crowther, D. (2016). Flawed self-assessment: Investigating self- and other-perception of second language speech. *Bilingualism: Language and Cognition*, *19*(1), 122–140. https://doi.org/10.1017/S1366728914000832

Truckenbrodt, H., Sandalo, F., & Abaurre, B. (2009). Elements of Brazilian Portuguese intonation. *Journal of Portuguese Linguistics*, *8*(1), 75–114. https://doi.org/10.5334/jpl.122

Tsunemoto, A., Trofimovich, P., Blanchet, J., Bertrand, J., & Kennedy, S. (2022). Effects of benchmarking and peer-assessment on French learners' self-assessments of accentedness, comprehensibility, and fluency. *Foreign Language Annals*, *55*(1), 135–154. https://doi.org/10.1111/flan.12571

Uchihara, T., & Clenton, J. (2020). Investigating the role of vocabulary size in second language speaking ability. *Language Teaching Research*, *24*(4), 540–556. https://doi.org/10.1177/1362168818799371

van Heuven, V. J., & Turk, A. (2021). Phonetic correlates of word and sentence stress. In C. Gussenhhoven & A. Chen (Eds.), *The Oxford handbook of language prosody* (pp. 150–165). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198832232.013.8

Wells, J. C. (2006). *English intonation: An introduction*. Cambridge University Press.

Winter, B. (2020). *Statistics for linguists: An introduction using R*. Routledge.

Wrembel, M. (2015). Metaphonological awareness in multilinguals: A case of L3 Polish. *Language Awareness*, *24*(1), 60–83. https://doi.org/10.1080/09658416.2014.890209

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Accessible Summary

Appendix S1. Stimuli Used in the Speech Elicitation Task and in the Assessment Tasks.

14679922, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/lang.12682 by University of Turku, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License