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

EMPIRICAL STUDY

Metacognition in Second Language Speech Perception and Production

Natalia Kartushina [a], David Soto [b,c] and Clara Martin [b,c]

[a]University of Oslo [b]Basque Center on Cognition, Brain, and Language [c]Ikerbasque, Basque Foundation for Science

Abstract: In this study, we assessed metacognition in nonnative language speech perception and production. Spanish novice learners of French identified and produced the French vowel contrast /ø/–/œ/ and, on each trial, rated their confidence in their responses. Participants' confidence in perception predicted their identification accuracy, suggesting that novice learners' metacognitive skills in nonnative speech perception are efficient at the onset of language learning. However, participants' confidence in production did not align with a fine-grained precision measure of their own production (indexed by Mahalanobis distance to the native French target-vowel space)

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

This project was supported by the Basque Government through the BERC 2022–2025 program and by the Spanish State Research Agency through BCBL Severo Ochoa excellence accreditation CEX2020-001010-S, by the Spanish Ministry of Economy and Competitiveness through project grants PID2020-113926GB-I00 to Clara Martin and PID2019-105494GB-I00 to David Soto, the Basque government (PIBA18_29 to Clara Martin), and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 819093 to Clara Martin). Natalia Kartushina was partly supported by the Research Council of Norway through its Centres of Excellence funding scheme (project number 223265). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Correspondence concerning this article should be addressed to: Natalia Kartushina, MuliLing, Institute for Linguistics and Scandinavian Studies, University of Oslo, 0315, Oslo, Norway. Email: natalia.kartushina@iln.uio.no. David Soto and Clara Martin, Basque Center on Cognition, Brain, and Language, Paseo Mikeletegi 69, San Sebastian 20009, Spain. Emails: c.martin@bcbl.eu; d.soto@bcbl.eu

The handling editor for this article was Theres Grüter.

DOI: 10.1111/lang.12549

nor with a categorical measure of production (in terms of being within/outside the native speakers' zone), indicating that metacognition in nonnative sound production is not yet efficient in novice learners. Overall, confidence ratings were similar and highly correlated between the perception and production tasks, but there was no association between the two domains in task performance or metacognitive ability. We discuss the ramifications of these findings for language learning theories and language teaching strategies as well as for the ongoing debate about the perception–production relationship.

Keywords metacognition; second language; speech; perception; production; self-confidence

Introduction

Metacognition is the process that allows individuals to reflect on and evaluate their own cognitive processes and performance in different tasks, including knowledge about the task and strategies needed to successfully achieve it (Flavell, 1979). Metacognitive evaluations come with a judgement of confidence that reflects the likelihood of a behavioral response to be correct (Maniscalco & Lau, 2012). In other words, metacognition is individuals' ability to be more confident in their correct relative to their incorrect responses.

Metacognition has been extensively studied in sensory perception (Fleming & Dolan, 2012; Jachs et al., 2015; Peters & Lau, 2015) and memory (Koriat, 2019; Koriat et al., 2006; Koriat & Shitzer-Reichert, 2002; Kuhlmann, 2019), demonstrating that people generally have good introspective access to the quality of their behavioral responses in perception and memory tasks. Self-reported confidence in such tasks has typically been higher for correct than for incorrect choices. These studies suggested that metacognition, through individuals' sense of confidence, is used to monitor their performance and may also promote the development of adaptive strategies to control behavior and learning (Nelson & Narens, 1994). Although metacognition has been extensively examined in attention and memory, the role of metacognition in language has remained to be evaluated systematically. In particular, there has been a gap in the understanding of the role of metacognition in foreign language learning, specifically as to how language learners monitor the correctness of their responses in linguistic tasks and whether the resulting metaknowledge can guide optimal learning strategies. Our study, therefore, investigated metacognition in language across two domains: speech sound perception and sound production.

Background Literature

Metacognition and L2 Perception and Production

In our study, we focused on novice language learners and the role of metacognition within nonnative sound perception and production. Foreign language learners typically display wide between- and within-speaker variability in their ability to accurately produce and perceive nonnative speech sounds. This variability is helpful when it comes to assessing metacognitive skills because it avoids ceiling effects likely observed in native production (Bradlow et al., 1997; Flege & Schmidt, 1995; Hanulíková et al., 2012; Nagle, 2018). Very little research has examined metacognition, that is, how individuals' confidence tracks the quality of behavioral responses, in perception and production of nonnative/second language (L2) speech sounds compared to larger linguistic units, for example, sentences, texts, and monologues. In earlier research, Yule, Damico, and Hoffman (1987) reported that L2 learners' self-monitoring ability to correctly identify L2 minimal word pairs was stronger in experienced learners who had undergone some formal learning compared to early-stage learners. In another study, Yule, Hoffman, and Damico (1987) investigated the effect of pronunciation-and-listening training on L2 learners' accuracy in phoneme identification using confidence ratings. The results showed considerable variation among participants in the size of the training effect on perceptual accuracy and also on the extent to which self-confidence ratings tracked response correctness (Yule, Damico, & Hoffman, 1987).

In the production domain, Dlaska and Krekeler (2008) studied the ability of advanced learners of German to monitor the quality of their speech production in comparison to the evaluations made by a native speaker. They found an overall correspondence between their learners' self-assessments and the native-speaker raters' assessments when the productions were indeed correct. However, learners had difficulties recognizing as incorrect a considerable number of productions, indicating that metacognitive monitoring of L2 speech production is not fully reliable. A recent study, also with German speakers, suggested that L2 learners, although aware of their accents in production, tend to overall judge their own production as less accented than those of other speakers of their language and to understand L2 words better when the words are produced in their own voice compared to the voice of other speakers (Eger & Reinisch, 2019). This was proposed to be due to more frequent exposure to their own productions, yielding less objectivity in their assessment (Mitterer et al., 2020). In line with this, a recent study showed that L2 speakers tended to misjudge their own L2 skills: Those who performed poorly overestimated their pronunciation skills, whereas those who

performed highly underestimated their skills (Trofimovich et al., 2016). This is reminiscent of the Dunning-Kruger effect in which inexperienced people think that they are better at cognitive or behavioral tasks than they truly are (Kruger & Dunning, 1999). Yet, experienced L2 speakers appear to have a more robust/clear overall estimation of their L2 production abilities, as their self-estimated pronunciation correlates strongly with the overall nativelikeness of their L2 production and with the accuracy in the production of specific L2 sounds, although this is to a lesser extent (Peperkamp & Bouchon, 2011).

Our review of a handful of studies on self-monitoring in L2 pronunciation and sound perception revealed inconsistent findings that could partly be attributed to differences in the unit of analysis (text vs. segment), in measures or methods used to assess L2 production (objective acoustic measures vs. perceptual ratings), and in L2 proficiency (novice vs. experienced). To the best of our knowledge, there has been a lack of research using objective measures of L2 production and, importantly, no studies examining metacognition in both L2 sound perception and production within the same experimental setting. Furthermore, prior work that we reviewed above did not address the trial-by-trial relationship between confidence-based self-monitoring and accuracy. In our study, we used objective measures to assess accuracy in nonnative speech sound production, and we assessed, within the same experimental setting, comparable levels of processing (i.e., phonemic) across the two language domains. We hypothesized that metacognition in novice learners would be reflected in a positive relationship between participants' trial-by-trial confidence ratings and their accuracy in L2 perception and production performance. We expected our study to be more sensitive than previous studies had been in pinpointing confidence-based metacognitive monitoring in novice learners.

Domain-Specific Versus Domain-General Metacognition

In addition to providing insights into metacognitive processes in language perception and production, examining metacognition in two language domains would also shed light on an unsettled debate on the nature of metacognition, that is, whether it is domain-specific or domain-general. A number of studies have suggested that metacognition operates at a domain-general level. Interindividual variability in the level of metacognitive sensitivity (i.e., how well confidence tracks trial-by-trial accuracy in a task) or in the associated neural markers of metacognition have been shown to correlate across different task domains (e.g., perception and memory), suggesting a domain-general metacognitive system (Ais et al., 2016; Baird et al., 2013; McCurdy et al., 2013; Morales et al., 2018; Samaha & Postle, 2017). The comparison of

metacognitive performance across language perception and production would provide knowledge on the organization/structure of metacognition in language. If we observed in our study that interindividual variation in metacognitive ability in perception was correlated with that of production, then we could conclude that both language domains are likely supported by a shared metacognitive system.

Perception–Production Interface

The final goal of our study was to examine the relationship between phonemic perception and production. This topic is of a particular relevance to research on nonnative/L2 processing. Late L2 learners often experience difficulties in perceiving and producing nonnative speech sounds, commonly revealed by their difficulties in discriminating L2 contrasts (e.g., *lock* vs. *rock* for Japanese learners of English) and their mispronunciation of L2 speech sounds (e.g., /i/ instead of /ɪ/ in *ship* for Spanish learners of English; e.g., Flege et al., 1997). Although the dominant theoretical perspectives have attributed L2 speakers' difficulties in production to a lack of accurate perceptual representations for L2 sounds and have predicted a tight relationship between the two domains (Best, 1995; Best & Tyler, 2007; Flege, 1995), research has provided no evidence for a consistent relationship between L2 speech sound perception and production (Bradlow et al., 1997; Flege, 1995; Flege & Eefting, 1987b; Hanulíková et al., 2012; Hattori & Iverson, 2010; Kartushina & Frauenfelder, 2014; Nagle, 2018; Nagle & Baese-Berk, 2021; Okuno & Hardison, 2016; Peperkamp & Bouchon, 2011; Sheldon & Strange, 1982). The strength of the relationship has been modulated by a number of variables, ranging from L2 experience and proficiency (Bohn & Flege, 1997; Jia et al., 2006; Rallo Fabra & Romero, 2012), to the level of linguistic processing explored (e.g., prelexical, phonological, lexical; see Bohn & Flege, 1997; Hao & de Jong, 2016; Melnik-Leroy et al., 2021; Peperkamp & Bouchon, 2011), L2 sound difficulty (e.g., similar to vs. distinct from native categories; e.g., Bohn & Flege, 1992; Evans & Alshangiti, 2018; Hao & de Jong, 2016; Levy, 2009; Levy & Law, 2010; Nagle, 2018), and L2 production accuracy measures (e.g., listener-based judgments vs. acoustic analyses; Evans & Alshangiti, 2018; Flege et al., 1999; Hattori & Iverson, 2010; Inceoglu, 2019). Therefore, for researchers to better understand the relationship between L2 perception and production, there must be strict and systematic control of participants' linguistic experience, the difficulty of the L2 sounds, and the tasks and measures for assessing L2 production. For our study, we attempted to control for these variables.

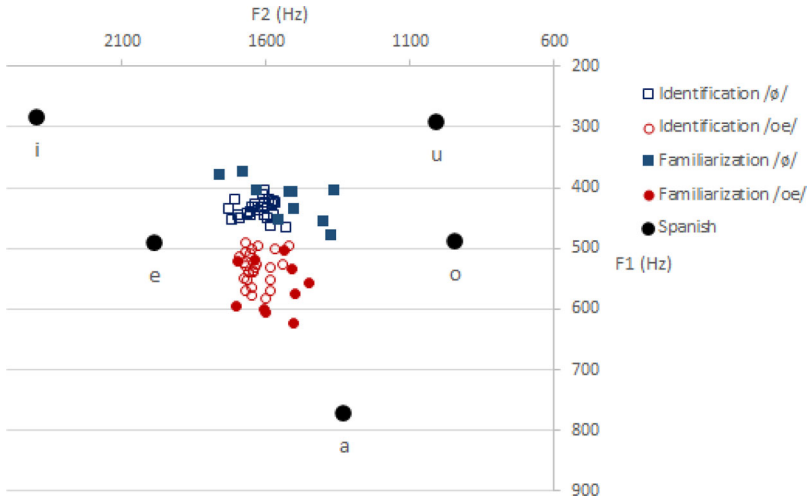

Figure 1 Distribution of the French vowel tokens used in the tasks and the Spanish norms for comparison (Chládková & Escudero, 2012).

The Present Study

To achieve the above-mentioned goals, we examined novice Spanish speakers' perception and production of the French vowel contrast /ø/–/œ/ and assessed, on a trial-by-trial basis, their self-confidence in their performance. Previous research (Kartushina & Frauenfelder, 2014) has shown that Spanish speakers experience difficulties in discriminating between these two vowels in both perception and production, suggesting that Spanish speakers perceptually assimilate both vowels to one new phoneme, dissimilar from Spanish categories, as Figure 1 shows.

Both Flege's (1995) speech learning model and Best's (1995) perceptual assimilation model agree that L2 sounds are processed as a function of their perceived similarity to close first language (L1) categories and to each other and predict that L2 learners' perception and production of the /ø/–/œ/ contrast will vary from poor to intermediate as a function of the learners' perception of the proximity of the /ø/–/œ/vowels to each other. In our study, the participants were familiarized with the target vowels /ø/–/œ/ and their respective labels through an auditory exposure to consonant–vowel (CV) words containing the target vowels. Then the participants performed, in a counterbalanced order, vowel identification and vowel reading tasks, including trial-by-trial confidence ratings of self-performance (see Figure 2). This experimental

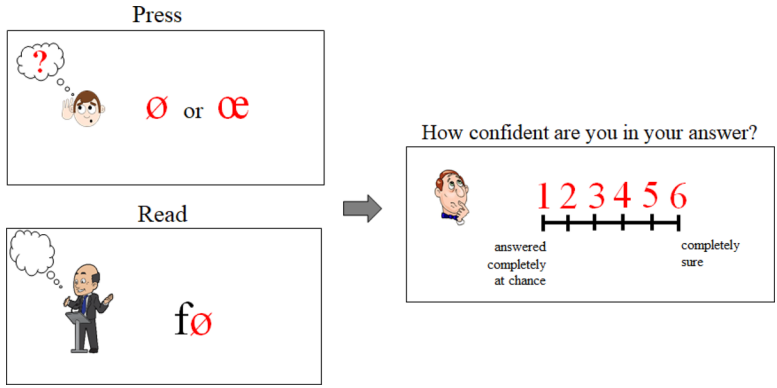

Figure 2 English translation of the visual aids displayed on the screen for the identification and reading tasks. In each of these tasks, on each trial, the participants had first to either press a button with the correct label or read a displayed consonant–vowel word and then to indicate the degree of confidence in their answer. The original material can be accessed through the project's OSF page (https://osf.io/usrdw) and via IRIS (see Kartushina et al., 2022c).

design allowed us to study the relationship between L2 speech sound perception and production both at the level of primary task performance and also at the level of metacognitive performance by systematically controlling several variables: L2 experience (all novice learners), level of processing assessed in both modalities (prelexical), no perceptual component in the production task, type of input during the exposure (identical familiarization for all participants), and type of L2 sounds (nonnative French vowels not present in the Spanish vowel inventory). In light of previous research and the above-mentioned specificities in our design, we expected to find a relationship between L2 speech sound perception and production. However, the relationship might be weak because we tested novice learners.

Method

Participants

We recruited 45 native Spanish speakers with no experience in French via the laboratory participant database. We excluded nine participants from the study: five participants due to technical reasons with the recording of their productions and four participants who failed to follow the instructions, as they did not provide confidence ratings on more than half of the trials. The final sample consisted of 36 speakers (M_{age} = 23 years, 22 females) who all

reported having normal hearing. All the participants provided consent prior to the experiment and received remuneration for taking part in the study. The project was approved by the Basque Center on Cognition, Brain, and Language Research Ethics Board.

Stimuli

Stimuli consisted of French word pairs featuring /ø/–/œ/ vowels.[1] A female native French speaker read aloud five word pairs featuring the target vowels in five different consonant contexts for which the consonants differed in manner and place of articulation, all of which exist in Spanish and are perceptually close to French: /dø/–/dœ/, /kø/–/kœ/, /fø/–/fœ/, /lø/–/lœ/, /pø/–/pœ/ (*deux–odeur* "two–smell," *queue–cœur* "line–heart," *feu–feur* "fire–nonword," *leu–leur* "leu [the currency in Romania]–their," *peu–peur* "little–fear"). The speaker repeated each word five times, yielding a total of 50 stimuli that were used in the vowel identification task. We used the same CV stimuli in the word reading task (dø–dœ, kø–kœ, fø–fœ, lø–lœ, pø–pœ). We have referred to this task as our nonword reading task because the participants were unfamiliar with these words.

In addition, five other female native French speakers read aloud a word pair featuring the target contrast in a different, sixth context using /s/: /sø/–/sœ/ (*ceux–sœur* "those–sister"). Each speaker repeated the word pair twice, yielding 20 stimuli that we used in the familiarization task. In our design, we wanted to encourage the participants to rely mostly on abstract phonological representations rather than to rely on a pure acoustic comparison of speech sounds. Yet, short words, as used in our study, can easily be compared at a purely acoustic level and result in ceiling performance. That is why, in the familiarization task, we opted for a single-consonant but multiple-talker context as the /s_/ frame was produced by five native French speakers, whereas in the identification task, we included contextual variability to maximize the participants' reliance on a more abstract representational level. An additional reason was to make the stimuli of the production and perception tasks uniform. Thus, by using untrained consonant contexts in the production task, we were able to examine spontaneous production as opposed to retrieval of a heard trace from the auditory memory.

We recorded all stimuli with a Marantz PMD670 portable recorder and sampled the recordings at 22.05 kHz directly to 16-bit stereo .wav files. We manually extracted the CV segments in Praat (Boersma & Weenink, 2020) and equalized them to 65 dB amplitude. We ramped the first 20 ms and last 50 ms. To examine the distribution of /ø/–/œ/ vowel exemplars in the acoustic vowel

space, we measured the average first two vowel formants, F1 and F2, using Praat default settings for female voice (maximum formants 5,500, five formants). We converted formant frequencies in Hz to Bark (Traunmüller, 1997) and adapted them for female voice following Bladon et al.'s (1984) formula: $(26.81 \times F/1960 + F) - 1.53$. Figure 1 illustrates the target French vowels compared to the Spanish norms (Chládková & Escudero, 2012). Importantly, the target French vowels had formant values comparable to the norms previously reported for native French female speakers (e.g., Georgeton et al., 2012).

Procedure

We assessed the native Spanish speakers' self-confidence in perception and production of the French contrast /ø/–/œ/ in a three-step procedure using the DMDX software (Forster & Forster, 2003). First, we familiarized participants with the target vowels: On each trial, they heard one of the CV syllables (/sø/ or /sœ/) and saw it written (in phonetic symbols) on the screen, with the target vowel highlighted in red. Their task was to learn the two novel vowel–label associations. There were 20 familiarization trials. Given that there were only two labels for the participants to learn and, on the basis of the results of our pilot tests, 20 trials seemed sufficient for the participants to learn the two sound–label associations. The results in the perception task demonstrated that, after the familiarization phase, the participants were above chance in demonstrating that they had learned the labels. Immediately after the familiarization phase, the participants proceeded in a counterbalanced order to the two remaining tasks: a phoneme identification task and a reading task. During both tasks, the participants wore Sennheiser Pro headphones equipped with a microphone that was used to deliver auditory stimuli and to record productions.

Identification Task

In the perceptual (identification) task, on each trial, the participants heard one of the target vowels embedded in a CV context (cf. the Stimuli section) and had to identify it by pressing the button labelled "ø" or "œ" that corresponded to the sound. Immediately after providing an answer, or at the end of a 3,000-ms timeout (if no answer was provided), the participants were asked to indicate how confident they were in the accuracy of their identification on a scale from 1 (*I answered completely at chance*) to 6 (*I am completely sure*). The participants were instructed to answer as quickly/intuitively as possible and within a timeout of 2,500 ms (see Figure 2 for an illustration). Upon providing an answer or at the end of the timeout, a 500-ms black screen was displayed before the next trial appeared. There were 200 trials in total, distributed across

four randomly presented blocks of 50 unique (unrepeated) trials. Therefore, each singular token was presented four times, and each vowel was displayed 100 times: 4 blocks × 5 consonant contexts × 5 speakers. The order of trials within each block was random. The participants were allowed to take a break between the blocks if they so desired.

Reading Task

In the reading task, the participants had to perform two actions similar to the perception (identification) task: They had to read a written CV nonword, consisting of a consonant and one of the novel vowel labels "ø" or "œ" and then to indicate the degree of self-confidence in the accuracy of their production/reading of the novel vowel (see Figure 2). On each trial, the participants saw a visual aid instructing them to read a nonword. This was followed 3,000 ms later by a self-confidence rating display that prompted the participants to report their degree of confidence on a 6-point scale as they had done in the perception task. The timeout, again, was set to 2,500 ms. Upon the participants' providing an answer or at the end of the timeout, a 500-ms black screen was displayed before the next trial appeared. There were 200 trials in total, distributed across four randomly presented blocks of 50 trials. Therefore, each vowel was produced 100 times (4 blocks × 5 consonant contexts × 5 repetitions).

Data Analysis

Identification Task

We coded the participants' accurate answers in the perceptual (identification) task as 1 and their incorrect answers as 0. The participants confidence rating scores ranged between 1 and 6. We removed trials with no answers from the analyses (1.07% of the data).

Reading Task

Individual audio recordings of the participants' productions in the CV nonword reading task (7,200 productions) underwent a four-stage processing. First, we denoised them. Second, we used a customized consonant-sensitive Matlab script to detect, for each word, vowel onset and offset, which we visually checked on a spectrogram and adjusted in Praat if necessary (this concerned 2.5% of data). Third, we computed the first two vowel formants, F1 and F2, over the whole vowel duration using the same automatized Praat script as the one that we had used to process native French speakers' productions (see the Stimuli section). We estimated 44 vowels manually because they were too short for the automatized script to process. We removed audio recordings

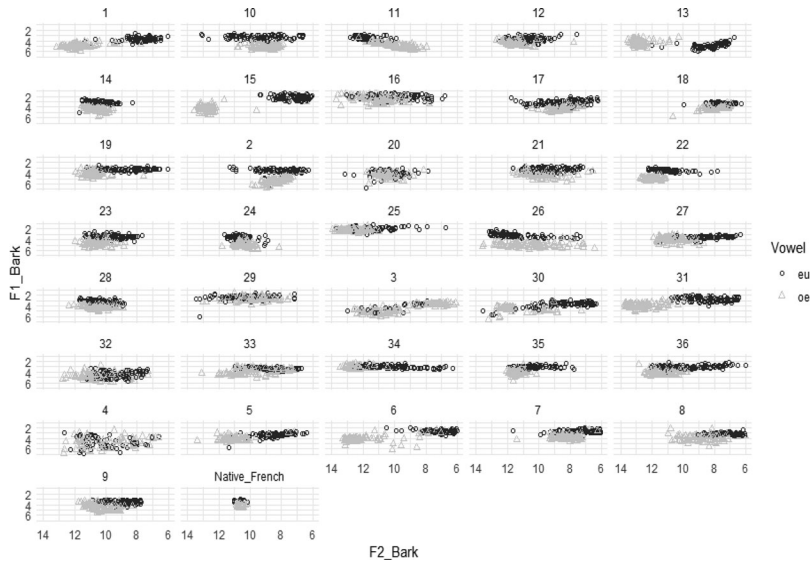

Figure 3 Participants' individual productions of the French /ø/ and /œ/ vowels (marked as "eu" and "oe" in the legend) in the acoustic F1/F2 space (in Bark), compared to native French speakers (Native_French).

containing coughs, sighs, laughs, or silence from the analyses (178, or 0.024%, of the recordings). Following the same procedure as we had used for the stimuli, we converted the formant frequencies in Hz to Bark. Figure 3 represents the participants' individual vowel productions in the acoustic F1/F2 space.

Finally, in order to assess the participants' accuracy in the production of novel French vowels in a manner similar to what has been done in previous studies (Franken et al., 2017; Kartushina et al., 2015, 2016; Kartushina & Martin, 2019), we computed, in a customized Matlab script, for each vowel token and participant, the Mahalanobis distance (or distance score [DS]) between this token and the target acoustic space, defined as the 20 vowel tokens produced by native French speakers that we had used for exposure in the familiarization phase. Mahalanobis distance (Mahalanobis, 1936) is a unitless, scale-invariant measure of distance in terms of standard deviation from a given point to a distribution that, by default, takes into account the noncircular shape of vowel categories and token distribution in the F1/F2 space. We computed the Mahalanobis distance for each individual vowel token. Therefore, for each participant, there were 100 distance scores for each vowel category. After removing nine outlying trials, the distance scores ranged between 0 and 15

($M = 4.06$, $SD = 2.46$). The 50 vowel tokens produced by native French speakers (and used in the identification task) were situated, on an average, at 1.2 distance score units to the target space (the most distant production was at 2.35 distance score units). Therefore, to dichotomize Spanish participants' accuracy in the production of French vowels, while accounting for the shape of the target vowel distribution, and to make it comparable to the binary accuracy measure used in the identification task, we categorized vowel tokens/productions with distance scores of less than 2.5 as within the native zone, whereas we categorized vowel tokens with distance scores greater than 2.5 as outside the native zone, according to the distribution of vowel tokens produced by the native French speakers recorded for this study that had fit all their vowels tokens into the 2.5 standard deviation area. The formant values for the target vowels recorded in our sample (i.e., defining the native zone) were comparable to the norms that have been reported elsewhere for native French female speakers (e.g., Georgeton et al., 2012). The mean proportion of vowel tokens produced by the Spanish learners within the native zone was 32% ($SD = 25\%$). Confidence ratings in the production task ranged between 1 and 6. We excluded from the analyses trials with no responses in the production task (0.94% of the trials).

Statistical Modeling

We conducted data processing and analyses using Python scripts (Version 3.7, https://www.python.org). We used Python's Scikit-learn to fit the linear regression and the logistic regression to test the relationship between confidence and accuracy in perception and production tasks (see Appendix S1 in the Supporting Information online for details).

We assessed metacognitive ability in the identification task also by computing meta-d'. Meta-d' provides an assessment of the efficacy with which participants' confidence ratings discriminate between correct and incorrect responses (i.e., Type-2 sensitivity). Meta-d' is a parametric estimation of participants' Type-2 sensitivity; it is computed by fitting a Type-1 signal detection theoretic model to the observed Type-2 performance data (Maniscalco & Lau, 2012) and estimating the Type-2 receiver operating characteristics (ROC) curves (i.e., based on the ratio of Type-2 hits—that indicate high confidence when participants' responses are correct—and Type-2 false alarms—that indicate high confidence when participants' responses are incorrect). Meta-d' presents clear advantages over other metrics used to assess metacognition such as the correlation between confidence and accuracy (see Fleming & Lau, 2014). In particular, meta-d' provides a measure of participants' metacognitive

sensitivity that is independent of individual biases in reporting high or low confidence. Also, both d' and meta-d' are on the same scale and can be compared directly, thereby allowing an assessment of participants' metacognitive ability regarding their level of performance. Readers can refer to Maniscalco and Lau (2012) for the specific computational procedure and the analytical scripts that they have developed meta-d'.

We conducted Bayesian analyses in JASP (JASP team, 2020) using the default priors. For the Bayesian correlation tests, we used a default beta prior width of 1. For the one-sample Bayesian t tests, we performed robustness checks and verified that Bayes factors[2] were robust to variations of the prior. We checked the normality assumption for the one-sample Student's t tests that we performed by using the Shapiro-Wilk test of normality, and when we detected a deviation from normality, we followed up on the one-sample t test with Wilcoxon signed-rank tests. We used a two-tailed alpha level of .05 to establish significance for all statistical tests. Experimental data, stimuli, experimental software, and analysis scripts are available on IRIS (Kartushina et al., 2022a, 2022b, 2022c) and via the OSF (https://osf.io/usrdw).

Results

Primary Task Performance in Nonnative Perception and Production

First, we examined the participants' primary task performance in their perception (vowel identification) and reading tasks. For that, we computed, for each participant, an averaged identification (i.e., accuracy) and production (i.e., Mahalanobis distance) score. A one-sample t test focusing on the participants' identification accuracy revealed significant above .50 chance performance, $M = 0.69$, $SD = 0.16$; 95% CI [0.64, 0.74], $t(35) = 7.21$, $p < .001$, $BF_{10} = 648{,}248.39$, Cohen's $d = 1.20$, 95% CI [0.77, 1.63]. Cohen's d indicated a large effect size. We obtained the same result with the perceptual sensitivity score—indexed by a bias-free, signal detection measure d' computed as z(hit rate) − z(false alarm rate), where a hit indicated that a participant responded "ø" when an /ø/ was actually presented, and a false alarm indicated that a participant responded "ø" when an /œ/ was present. Perceptual sensitivity was clearly above 0 chance, $M = 1.09$; $SD = 0.98$; 95% CI [0.76, 1.42], $t(35) = 6.69$, $p < .0001$, $BF_{10} = 34.36$, Cohen's $d = 1.12$, 95% CI [0.69, 1.53]. Cohen's d again indicated a large effect size. We then assessed with a Spearman rank-order correlation whether individual perceptual accuracy was associated with individual performance in production, production being indexed by Mahalanobis distance scores. As the analysis revealed, there was no evidence

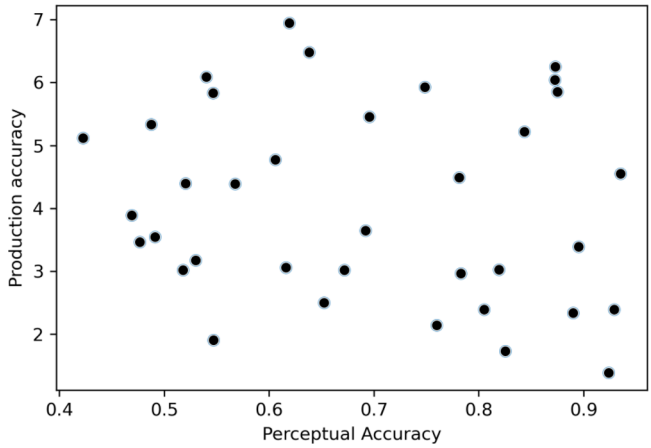

Figure 4 Interindividual Spearman rank-order correlations between accuracy in nonnative vowel identification and vowel production (nonword reading) in Mahalanobis distance.

of a positive correlation, $r_s = -.17$, $p = .31$, $BF_{10} = 0.34$, providing moderate evidence for the null hypothesis (see Figure 4), suggesting that nonnative speech sound perception and production might rely on distinct sound representations in novice learners.

Metacognitive Performance in Nonnative Sound Identification and Production

Second, we assessed the participants' metacognitive performance (i.e., the ability to endorse higher confidence ratings for correct relative to incorrect responses) in both perception and production tasks. In the perception task, we used a logistic regression to examine how accuracy (0 vs. 1) related to trial-by-trial confidence ratings. As a dependent metric, we used the area under the ROC curve of the logistic regression. The ROC is a sensitive, nonparametric bias-free measure of predictive performance in binary classification, with .50 being the theoretical chance level. The ROC represents the ratio of the true positive rate (i.e., the regression predicts "correct" given that the trial is correct) against the false positive rate (i.e., the regression predicts "correct" given that the trial is incorrect). We computed this separately for each participant (ROC scores ranged from 0 to 1, with .50 indicating chance level). Figure 5 shows the distribution of the ROC scores of the logistic regression across participants. These were significantly higher than

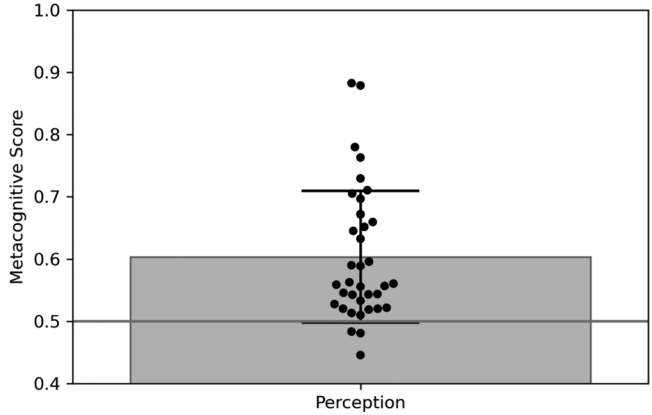

Figure 5 Metacognitive performance in vowel identification task indexed by the receiver operating characteristics scores of a logistic regression analysis relating confidence ratings to the identification accuracy. The bars represent the 95% confidence intervals around the mean. The grey filled bar plots the mean. Dots represent individual data points.

.50 chance, $M = .60$, $SD = .11$, 95% CI [.57, .64], $t(35) = 5.79$, $p < .001$, $BF_{10} = 12{,}182.70$, Cohen's $d = 0.97$, 95% CI [0.56, 1.36], indicating that confidence ratings significantly predicted perception accuracy, and the effect size was large. Given that we had detected deviation from normality in the ROC scores, we verified the result with a Wilcoxon signed-rank test, $Z = 644$, $p < .001$, Cohen's $d = 0.93$, 95% CI [0.87, 0.97]. The effect size remained large.

We also confirmed this result using measures of metacognitive sensitivity based on signal detection theory (i.e., meta-d'). Meta-d' was also above 0 chance, $M = 0.74$, $SD = 0.94$, 95% CI [0.42, 1.06], $t(35) = 4.70$, $p < .001$, $BF_{10} = 583.60$, Cohen's $d = 0.78$, 95% CI [0.41, 1.15]. The effect size was medium. Given that we had detected a deviation from normality in the meta-d' scores, we verified the result with a Wilcoxon signed-rank test, $Z = 579$, $p < .001$, Cohen's $d = 0.74$, 95% CI [0.52, 0.87]. The effect size was also medium in this analysis.

Since production accuracy was a continuous response, we were not able to use the meta-d' metric to compare metacognitive performance in production and perception. To analyze metacognition in the production task, we used a linear regression model to predict the precision of the production (distance scores) for each participant across trials based on the trial-wise confidence ratings. Figure 6 depicts the distribution of the regression coefficients, which

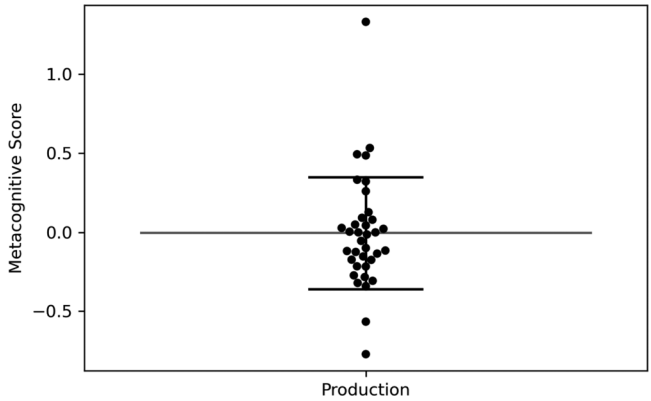

Figure 6 Metacognitive performance in nonword reading indexed by a linear regression relating confidence ratings to production performance. The bars represent the 95% confidence intervals around the mean. Dots represent individual data points.

were low, on average, and not significantly different from 0, $M = -0.01$, $SD = 0.36$, 95% CI [−0.13, 0.01], $t(35) = -0.12$, $p = .908$, $BF_{10} = 0.18$, Cohen's $d = -0.02$, 95% CI [−0.35, 0.31]. Given that we had detected a deviation from normality in the production scores, we verified the result with a Wilcoxon signed-rank test, $Z = 248$, $p = .402$, Cohen's $d = -0.16$, 95% CI [−0.50, 0.21].

This result indicated that the participants did not display a fine-grained metacognitive insight about the quality of their production, as indicated by lack of a relationship between confidence ratings and the acoustic distance from their productions to the target native vowel. Measuring metacognition using regression models has potential drawbacks as estimations may be influenced by the confidence bias, namely, the tendency of individuals to be over- or under-confident (Fleming & Lau, 2014), but they, however, reliably indicate whether confidence ratings track the correctness of behavioral responses (Rahnev et al., 2020). Individual confidence biases were similar and correlated across the perception and production tasks, suggesting that they could not account for the results of the regression analyses in metacognition that we have reported.

In a separate post hoc analysis, we addressed whether the participants had a less refined, binary sense of their accuracy in terms of the nativeness of their production (nativelike/not nativelike), rather than a fine-grained continuous representation of their production accuracy. Therefore, as we mentioned previously, we dichotomized the responses in the production task as a function

of their location within or outside the French native vowel space and assigned each nonnative production to either of the two categories: within or outside the native zone. Initially, as in the analysis of the perception data, we used a logistic regression model to assess whether confidence ratings predicted the production accuracy (within vs. outside the native space). Although the results appeared to indicate an above-chance relationship, a check of the raw confidence ratings revealed that confidence was not related to production accuracy. Specifically, the participants' confidence level was not significantly different between the productions within and outside the native space, $M_{\text{within}} = 4.39$, $SD = 0.84$, $M_{\text{outside}} = 4.46$, $SD = 0.86$, 95% CI [−0.20, 0.07], $t(35) = -0.98$, $p = .330$, $BF_{10} = 0.28$, Cohen's $d = -0.16$, 95% CI [−0.49, 0.17]. Given that a deviation from normality was detected in the confidence levels, we verified the result by means of a Wilcoxon signed-rank test, which revealed a similar nonsignificant effect, $Z = 268$, $p = .830$, Cohen's $d = -0.05$, 95% CI [−0.41, 0.33]. Therefore, confidence did not track participants' accuracy in production.

Role of Metacognitive Ability in Sound Perception and Production

We assessed, using the Spearman rank-order correlation, the relationship between interindividual variation in metacognitive sensitivity and nonnative speech sound perception and production. The results did not reveal a relationship between metacognitive sensitivity and perception or production when using either the distance, $r_s = .07$, $p = .684$, $BF_{10} = 0.24$ (see Figure 7). We also assessed whether the participants' feeling of confidence was correlated across the two language domains; these analyses can be found in Appendix S2 in the Supporting Information online.

Discussion

In our study, we developed a paradigm to investigate, for the first time in a single study and within the same participants, metacognition in nonnative language (phonemic) perception and production. To do this, we familiarized L1-Spanish speakers with the novel French vowel contrast /ø/–/œ/ and then assessed their confidence in both a vowel identification and a nonword reading task, on a trial-by-trial basis. Our main goals were: (a) to examine metacognition in two different but related language domains — speech sound perception and production; (b) to examine the relationship between metacognition in speech perception and production; and (c) to address the relationship between phonemic perception and production in novice language learners.

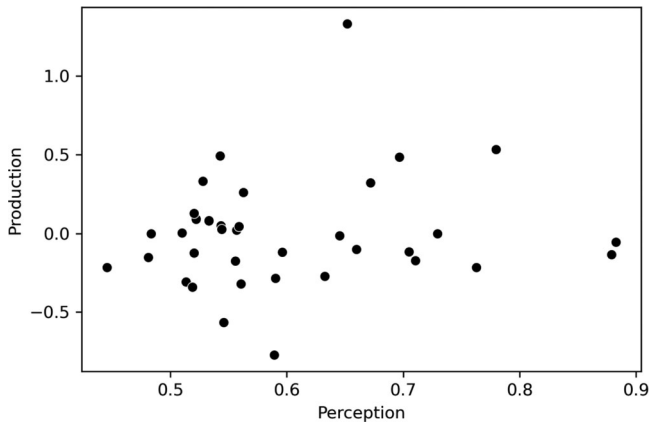

Figure 7 Absence of a relationship between metacognitive ability in vowel perception (identification) and in vowel production (nonword reading). Metacognitive sensitivity in perception was indexed by the receiver operating characteristics scores of the logistic regression predicting accuracy by confidence. Metacognitive sensitivity in production was indexed by the linear regression coefficient relating confidence to accuracy.

Perception and Production of L2 Speech Sounds

Regarding the primary task performance, in line with previous research (Kartushina & Frauenfelder, 2014), Spanish speakers demonstrated difficulties in their perception and production of the French /ø/–/œ/ contrast. Although, overall, the participants identified the target vowels above chance, their perception accuracy was rather moderate (69% correct), and the average speech production distance was around four standard deviations from the mean of the target native (French) space, with only 32% of productions realized within the native target vowel zone. This indicated, in line with Best's (1995) perceptual assimilation model and Flege's (1995) speech learning model, that similar nonnative speech sounds having no close counterparts in the native (vowel) space, assimilate to one uncategorized novel sound, and are likely to be misperceived and/or mispronounced until listeners can discern the phonetic difference between them. Our results revealed considerable interspeaker variability in perceiving and producing the novel difficult nonnative contrast, demonstrating, similar to the findings of previous research, the role of individual-specific variables in L2 speech sound processing (Bradlow et al., 1997; Flege & Schmidt, 1995; Hanulíková et al., 2012; Nagle, 2018).

We found no relationship between individual accuracy in speech perception (vowel identification) and production (nonword reading) in novice learners,

suggesting the absence of a link between the two domains, at least when individuals are learning novel perceptually similar L2 sounds distinct from native categories. This result is consistent with previous research showing that, while experienced L2 speakers show a moderate-to-strong relationship between L2 perception and production (Baker & Trofimovich, 2006; Bettoni-Techio et al., 2007; Flege, 1993, 1999; Flege et al., 1997, 1999; Flege & Eefting, 1987a; Flege & Schmidt, 1995; Hattori & Iverson, 2010; Jia et al., 2006; Kluge et al., 2007; Zhang & Peng, 2017), speakers with less experience or novice learners show only weak or no relationship at all (Jia et al., 2006; Kartushina et al., 2015; Kartushina & Frauenfelder, 2014; Li et al., 2019). This may indicate that the two domains align with experience (Flege & Schmidt, 1995), as the perception–production link varies across time (Nagle & Baese-Berk, 2021) and has been shown to be stronger in experienced L2 speakers (Rallo Fabra & Romero, 2012). Yet, even experienced L2 speakers might show a dissociation between the two domains when the tasks used to assess L2 perception and production tap into different processing levels (e.g., Hao & de Jong, 2016; Melnik-Leroy et al., 2021; Peperkamp & Bouchon, 2011). For instance, while no relationship has been reported between L2 speech perception and production at the prelexical, acoustic–phonetic level when acoustic measures for L2 production accuracy are used (Hattori & Iverson, 2010; Schertz et al., 2015), a moderate-to-strong relationship has been revealed at the phonological/lexical level when researchers use more global perceptual (e.g., subjective) measures of perception and production accuracy (e.g., through listener-based judgments, see Evans & Alshangiti, 2018; Flege et al., 1999; Inceoglu, 2019) or tapping into similar levels of processing (Bohn & Flege, 1997; Flege, 1993; Flege et al., 1997, 1999; Flege & Schmidt, 1995; Melnik et al., 2021; Nagle, 2018). In our study, we carefully controlled for input/stimuli in exposure and during the tasks (we used unfamiliar phonetic symbols "ø" or "œ" for novel nonnative vowels), aiming at assessing nonnative perception and production at a similar prelexical, that is, phonemic level. We did not use a sound repetition task to avoid the participants' relying on acoustic processing only (i.e., imitation), and we used nonwords to avoid lexical effects. Nevertheless, the results revealed no association between the two domains in novice learners, suggesting that either the relationship is not yet stable at the onset of phonetic learning or that it might be lagging as recent research has suggested. Longitudinal studies assessing the development of the perception–production relationship are needed to provide further support for this hypothesis (see Nagle & Baese-Berk, 2021, for a review).

Another variable that might have contributed to the lack of a correlation of L2 sound perception with production was our use of acoustic measures to assess L2 production. In our study, we opted for an acoustic analysis of nonnative vowel production, instead of more subjective accuracy ratings that have typically been used to assess overall word production, including individual segments, suprasegmental information, and coarticulation, and can be subject to bias due to raters' familiarity with the learners' language/accent (Winke et al., 2013), raters' background and attitudes (Kang et al., 2019), and learners' fluency (Duijm et al., 2018). Although there seems to be a general relationship between acoustic and listener-based measures of L2 production (e.g., see Flege et al., 1997; Wang et al., 2003; although see Delvaux et al., 2013, showing only partial relationship)—providing grounds for using either of them as an informative pronunciation measure—acoustic analyses provide a finer-grained (continuous) measure of accuracy. This might not be captured by a dichotomized measure of perceptual accuracy (as used in our study) compared to a two-forced-choice identification task over a continuum, for example, from the French /ø/ to /œ/. For instance, previous research in L2 learning revealed no relationship (at the prelexical level) between L2 speech perception and production when acoustic measures of L2 production were used (Bohn & Flege, 1997; Schertz et al., 2015) compared to when listener-based measures were used, including those that assess L2 pronunciation beyond segmental accuracy, that is, suprasegmental, coarticulation, and the like (Evans & Alshangiti, 2018; Flege et al., 1999; Inceoglu, 2019). In contrast, a study by Hattori and Iverson (2010) showed no relationship between the production accuracy of the relevant acoustic cues and perceptual sensitivity to these cues. These inconsistencies might be partially attributed to the type of acoustic cues used to assess L2 production. As Nagle and Baese-Berk (2021) suggested in a recent review paper, acoustic features that distinguish L2 sounds in native speakers' speech might differ from those that L2 learners use when attempting to differentiate L2 contrasting sounds in production, hence more complex acoustic analyses, taking into account a multitude of features, might be needed to fully describe L2 pronunciation accuracy and the acoustic features that L2 learners use to differentiate L2 contrasting sounds (e.g., linear discriminant analysis, see Mairano et al., 2019). In a recent study, Song and Eckman (2021) found a relationship between L2 speech perception and production when the acoustic analysis of L2 speech included the same set of cues that L1 speakers use to distinguish the target vowel contrast (duration, F1, and F2, although with use of a binary response, such as a cue is present vs. absent), but not when the production was assessed via listener-based judgements, suggesting that

fine-grained acoustic analyses of L2 production might be better suited for assessing L2 segmental pronunciation compared to raters' judgements.

In our study, to assess L2 speakers' production, we used only F1 and F2, as these are the main cues that native French speakers use to distinguish between the two front rounded vowels. However, it is not impossible that native Spanish speakers' rounding of these vowels (revealed by the third formant, F3) differed from that implemented by native French speakers, and adding this cue in the accuracy measure would have increased the strength of the perception–production relationship (see Llompart & Reinisch, 2018, for work on the role of acoustic cues in L2 perception). To examine this possibility and the role of F3 in the production of this novel contrast, we ran a series of analyses (available in the project's OSF profile at https://osf.io/usrdw, see also Kartushina et al., 2022a), that revealed: (a) no difference in F3 between Spanish and native French groups, nor interaction between group and vowel, (b) no correlation between the F3-distance to the norms (participants' F3 and native French speakers' F3) and the confidence rating, and (c) no correlation between the F3-distance to the norms and participants' identification accuracy for either vowel. These additional analyses suggested that, similar to native speakers of French, Spanish participants did not use F3, a roundness cue, to distinguish between the two French vowels and that the accuracy in F3 (alone) was not related to the participants' contrast perception. Future research needs to examine fine-grained representations underlying L2 sound perception and production across different learner profiles and tasks using more exhaustive measures of production, as, for instance, measures over the three acoustic cues F1, F2, and F3, use of a 3D distance metric, and measures of perception accuracy.

Metacognition in L2 Speech Sound Perception and Production

Regarding metacognition, we observed that the participants had insights into their accuracy in nonnative sound perception: their confidence ratings were associated with their accuracy in nonnative vowel identification. This result, stemming from a rich trial-by-trial analysis of participants' performance (200 trials per participant), is in line with the findings of previous L2 learning research where self-confidence has been rated according to individuals' overall task performance. For instance, Yule, Hoffman, and Damico (1987) showed that improvements in nonnative sound perception following training were accompanied by improvements in L2-learners' self-confidence. Yet, in learners with low L2 levels, Yule, Damico, and Hoffman (1987) revealed that overall improvements in L2 identification accuracy were not accompanied by improvements in self-monitoring, suggesting that learners at low L2 levels (typically at

the outset of learning) might be most focused on improving their identification skills and unaware of their progress. However, in contrast with our study, Yule, Damiclo, and Hoffman did not assess the relationship between confidence and accuracy on a trial-by-trial basis but only self-monitoring at the beginning and at the end of training. Our results in novice learners, on the other hand, are based on a more sensitive, trial-by-trial analysis and indicate that metacognitive ability in nonnative speech sound perception can be efficient from the onset of learning. Though future research is still needed, we propose, tentatively, that this metacognitive ability can guide foreign language learners' acquisition of difficult L2 sounds.

The participants' confidence ratings in the production task, on the other hand, did not predict the quality of their production, which was indexed by a continuous acoustic measure of distance from the participants' production to the target native space, nor when the participants' production quality was captured through a categorical measure of nativeness in terms of their production being within or outside the native vowel zone. These results suggest that metacognition in nonnative speech sound production in novice learners is less efficient compared to metacognition in nonnative speech sound perception.

Previous research that used a nativeness criterion to measure L2 (segmental) production accuracy reported robust overall estimation of L2 production abilities: Participants' self-estimated pronunciation correlated strongly with the overall nativelikeness of their L2 production and, to a lesser extent, with the accuracy in the production of specific L2 sounds (Peperkamp & Bouchon, 2011). Similarly, Dlaska and Krekeler (2008) reported a moderate alignment between experienced L2 speakers' self-assessments and raters' assessments of their accuracy. However, L2 speakers had difficulties in recognizing a considerable number of productions being mispronounced, indicating that metacognition in L2 speech production might still be relatively modest, even in experienced L2 speakers. When assessing broader linguistic levels (self-assessment of sentence production, discourse), L2 speakers mostly misjudge their production accuracy (Trofimovich et al., 2016) and might not be aware of their overall comprehensibility (Strachan et al., 2019), suggesting that multilevel monitoring, involving microlinguistic (e.g., sounds) and macrolinguistic structures, is a complex cognitive skill that can be beyond the reach of even experienced L2 speakers. Alternatively, better acceptability of self-produced accented speech (i.e., failure to detect mispronunciations) and better comprehension of words produced by an individual's own voice can be attributed to the familiarity effect, suggesting that individuals' exposure to their own accented productions can induce adaptation that, in turn, can hinder changes in L2 experienced

speakers' pronunciation (Eger & Reinisch, 2019; Mitterer et al., 2020) and might encourage them to seek input from native (unfamiliar with the accent) speakers (Carey et al., 2011). Similar familiarity effects have been reported in native raters' assessment of L2 pronunciation, suggesting that exposure to a foreign accent might facilitate adaptation for this specific accent but hinder the reliability of the assessment (Carey et al., 2011).

Successful L2 speech sound learning frequently involves (a) distinguishing cross-linguistic differences between native and nonnative speech sounds, which supports the establishment of novel L2 categories, and (b) discerning differences between similar nonnative speech sounds, which is necessary for establishing contrasting target-language-like sound categories used in minimal word pairs like *ship–sheep* in English (Flege & Bohn, 2021). In our study, we could not disentangle these two processes. Presumably, those Spanish listeners who were able to correctly identify the French contrast and confidently report it, discerned the cross-linguistic differences between these French sounds and similar Spanish sounds (e.g., back rounded mid /o/ and, to a lesser extent, front unrounded mid /e/). However, it should be noted that French front rounded vowels do not systematically map onto one specific Spanish category, but rather onto a new category (in the speech learning model terms) with the two vowels assimilating to one uncategorized vowel (in the perceptual assimilation model terms); hence, native Spanish speakers' ability to correctly assess their perception and production of the French contrast can also reflect their ability to distinguish the two nonnative vowels. The results of our study suggest that metacognition could guide L2 sound learning by providing learners with self-confidence and support in the learning process, whether that is to distinguish cross-linguistic differences between L2 and similar L1 speech sounds or between similar L2 speech sounds, in perception and production or both. Future research needs to address what types of L2 sound learning are facilitated by greater metacognitive awareness and how metacognition can be incorporated into current L2 learning models.

Metacognition in L2 Speech Sound Perception and Production: Domain General or Domain Specific?

Finally, the results of the regression analyses revealed no evidence for an association between the participants' metacognitive ability in speech sound perception and production. Therefore, our results do not suggest that metacognition in nonnative language learners is mediated by a domain-general system operating across the two domains of phonemic perception and production, similar to what has been suggested for other cognitive processes (see Rouault et al.,

2018, for a meta-analysis in the perceptual and memory domains suggesting that both domain-general and domain-specific monitoring systems might coexist). However, the lack of alignment between the metacognitive processing in (nonnative) speech sound perception and production can also be attributed to differences in the underlying processes and/or to immaturity or instability of the phonemic representations for the newly learned nonnative speech sounds.

While self-assessment of individuals' own production, on the one hand, relies on a goodness of fit between the motor, somatosensory, and auditory consequences of the sound pronunciation (or output) and the phonemic representation for this specific sound, or the auditory target (Guenther, 2006), self-assessment of individuals' own perception, on the other hand, relies on a goodness of fit between the auditory input (incoming sound) and the phonemic representation for this specific sound, the auditory target. Previous research has suggested that when speaking a nonnative language, L2 learners might have less efficient postarticulatory sensory monitoring (Simmonds et al., 2011) that is used to adjust and correct sound articulation online to match the auditory target (Guenther, 2006). This deficiency can disadvantage L2 learners' assessment of their own production, leading to a decrease in the quality and/or quantity of available information for them to assess segmental production. Production monitoring over larger linguistic units (words, sentences) might be less focused on individual sounds but rather take into account other relevant global features, for example, fluency, overall accentedness, and suprasegmental cues, among others, that provide richer input for the assessment. This, however, might compromise the assessment (in particular in novice learners) as multilevel monitoring is required (Trofimovich et al., 2016). Language experience can contribute to improvements in metacognitive processing by either strengthening weak sensory monitoring (for instance, via intensive experience in L2 speech sound production, cf. Simmonds, 2015) and/or tuning the auditory targets. Additional work is needed to further tackle this issue.

Though there was no association between metacognitive ability across the two domains, individual confidence ratings for the perception task significantly correlated with those of the production task. This result indicated that the participants felt similarly confident and experienced similar task difficulty related to their performance across the two language tasks (nonnative sound perception and production), even though the actual metacognitive performance did not correlate across them. Importantly, the correlation of self-confidence in perception with production suggests that the lack of an association in metacognitive performance across the two tasks was not due to differences in the participants' tendency to report high/low confidence. The absence of a

correlation between metacognition in language perception and production is consistent with the results of a prior meta-analysis by Rouault et al. (2018) and indicates that metacognitive associations across different domains of perception and memory, for example, are likely to be low and highly variable across people. However, to the best of our knowledge, our study was the first to address metacognition across the domains of (non native) language perception and production, and additional work is needed to make further assessment of variables that may constrain metacognitive insight across these domains and across the domains of perception and action more broadly (e.g., paradigms jointly assessing confidence in both sensory and motor, nonlinguistic signals).

Implications

Our work has an important pedagogical implication. Previous research has already shown that individuals can improve their ability to assess their own accuracy in L2 speech sound production and perception (Yule, Damico, and Hoffman, 1987; Yule, Hoffman, and Damico, 1987), yet little is known about whether this ability can be trained. In light of our results suggesting that self-assessment (reliance on self-confidence) can successfully guide L2 learning, training individuals to become more accurate at assessing their own perception and production can lead to better learning outcomes and facilitate L2 learning. This is important as listeners' assessment of L2 speakers' production can be unstable and vary as a function of the listeners' age and native language (Saito et al., 2019) and can be prone to social bias (Reid et al., 2019) and task effects (Crowther et al., 2018). The hypothesis that training metacognition can increase the efficiency of L2 learning is further supported by research on the use of metacognition in classroom teaching, in a broader sense, showing that metacognitive instruction can enhance the effectiveness of corrective feedback (Sato & Loewen, 2018). Our results demonstrating a lack of efficient metacognitive skills in novice learners' production suggest that novice learners' L2 sound representations operate at a relatively coarse level, which can limit the use of these representations in a fine-grained way to tune their production at the onset of learning. Further learning can rely more heavily on L2 teachers' feedback, which has been proven to be equally effective whether teachers of L2 pronunciation are native or nonnative (Henderson et al., 2012; Levis et al., 2016). Teachers' feedback can also be used with students with low metacognitive skills, which might be related to anxiety, among other variables (cf. MacIntyre et al., 1997).

In sum, our findings suggest that L2 novice learners have an accurate overall estimation of their L2 phonemic skills: They accurately self-monitor

outcomes in their L2 speech perception, yet the metacognitive system seems to weakly monitor the precision of L2 production, which is likely related to unstable or imprecise representations for novel nonnative speech sounds. These results are encouraging and valuable for foreign-language teaching research, as they suggest that a brief auditory exposure to nonnative speech sounds might suffice to guide perceptual learning and further pronunciation tuning in novice adult learners. Future research needs to address the development of metacognitive monitoring in L2 processing as a function of experience and proficiency, as recent research has suggested that individuals' familiarity with their own accented voice leads to accent adaptation/acceptability that might prevent pronunciation improvement in L2 experienced learners (Mitterer et al., 2020).

Conclusion

In conclusion, the results of the current study shed new light on the mechanisms involved in nonnative speech perception and production and on the role of metacognition in L2 language processing. Previous research had already highlighted the role of individual differences in L2 learning (Golestani & Zatorre, 2009; Inceoglu, 2019; Schertz et al., 2015) and of metacognition in L2 classroom based learning (Sato & Loewen, 2018); our study has contributed to this line of research and provided evidence that individual differences in metacognition (i.e., individuals' ability to assess their own L2 perception and production) can also account for variability in L2 learning and can be used as an anchor for facilitating L2 speech sound acquisition, and potentially for enhancing learning outcomes with training. We believe that future research along the lines of our study has the potential to inform the development of novel and more accurate strategies of self-assessment during language learning, to encourage the development of novel approaches to L2 training, and to account for differences in metacognition in L2 models and L2 teaching.

Final revised version accepted 10 August 2022

Notes

1 In order to have identical CV structure for both vowels, we asked native French speakers to pronounce the final consonant in CœC words (e.g., *coeur*, *odeur*, *feur**, *leur*, *peur*) silently.

2 The Bayes factor (BF_{10}) is used to quantify the level of evidence in the data for the alternative/null hypotheses. Bayes factors between (a) 1–3, (b) 3–10, (c) 10–30, and (d) 30–100 are respectively considered (a) anecdotal, (b) moderate, (c) strong, and (d) very strong evidence for the alternative hypothesis. Inversely, Bayes factors between (a) 1–0.33, (b) 0.33–0.1, (c) 0.1– 0.03, and (d) 0.03–0.01 are respectively

considered (a) anecdotal, (b) moderate, (c) strong, and (d) very strong evidence for the null hypothesis (Quintana & Williams, 2018).

Open Research Badges

This article has earned Open Data and Open Materials badges for making publicly available the digitally-shareable data and the components of the research methods needed to reproduce the reported procedure and results. All data and materials that the authors have used and have the right to share are available at https://osf.io/usrdw/ and http://www.iris-database.org. All proprietary materials have been precisely identified in the manuscript.

References

Ais, J., Zylberberg, A., Barttfeld, P., & Sigman, M. (2016). Individual consistency in the accuracy and distribution of confidence judgments. *Cognition*, *146*, 377–386. https://doi.org/10.1016/j.cognition.2015.10.006

Baird, B., Smallwood, J., Gorgolewski, K. J., & Margulies, D. S. (2013). Medial and lateral networks in anterior prefrontal cortex support metacognitive ability for memory and perception. *Journal of Neuroscience*, *33*(42), 16657–16665. https://doi.org/10.1523/JNEUROSCI.0786-13.2013

Baker, W., & Trofimovich, P. (2006). Perceptual paths to accurate production of L2 vowels: The role of individual differences. *International Review of Applied Linguistics in Language Teaching*, *44*(3), 231–250. https://doi.org/10.1515/IRAL.2006.010

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues* (pp. 171–204). York Press.

Best, C. T., & Tyler, M. D. (2007). Non-native and second-language speech perception: Commonalities and complementarities. In M. J. Munro & O.-S. Bohn (Eds.), *Second language speech learning: The role of language experience in speech perception and production* (pp. 13–34). John Benjamins.

Bettoni-Techio, M., Rauber, A. S., & Koerich, R. D. (2007). Perception and production of word-final alveolar stops by Brazilian Portuguese learners of English. *Proceedings of Interspeech 2007*, 2293–2296. https://doi.org/10.21437/Interspeech.2007-622

Bladon, R. A. W., Henton, C. G., & Pickering, J. B. (1984). Towards an auditory theory of speaker normalization. *Language & Communication*, *4*(1), 59–69. https://doi.org/10.1016/0271-5309(84)90019-3

Boersma, P., & Weenink, D. (2020). *Praat: Doing phonetics by computer* (Version 6.1.13) [Computer software]. http://www.praat.org

Bohn, O.-S., & Flege, J. (1992). The production of new and similar vowels by adult German learners of English. *Studies in Second Language Acquisition*, *14*(2), 131–158. https://doi.org/10.1017/S0272263100010792

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

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

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, *28*(2), 201–219. https://doi.org/10.1177/0265532210393704

Chládková, K., & Escudero, P. (2012). Comparing vowel perception and production in Spanish and Portuguese: European versus Latin American dialects. *The Journal of the Acoustical Society of America*, *131*(2), EL119–EL125. https://doi.org/10.1121/1.3674991

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

Delvaux, V., Huet, K., Piccaluga, M., & Harmegnies, B. (2013). Production training in second language acquisition: A comparison between objective measures and subjective judgments. *Proceedings of Interspeech 2013*, 2375–2379. https://doi.org/10.21437/Interspeech.2013-554

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

Duijm, K., Schoonen, R., & Hulstijn, J. H. (2018). Professional and non-professional raters' responsiveness to fluency and accuracy in L2 speech: An experimental approach. *Language Testing*, *35*(4), 501–527. https://doi.org/10.1177/0265532217712553

Eger, N. A., & Reinisch, E. (2019). The impact of one's own voice and production skills on word recognition in a second language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(3), 552–571. https://doi.org/10.1037/xlm0000599

Evans, B. G., & Alshangiti, W. (2018). The perception and production of British English vowels and consonants by Arabic learners of English. *Journal of Phonetics*, *68*, 15–31. https://doi.org/10.1016/j.wocn.2018.01.002

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Flege, J. (1993). Production and perception of a novel, second-language phonetic contrast. *The Journal of the Acoustical Society of America*, *93*(3), 1589–1608. https://doi.org/10.1121/1.406818

Flege, J. (1995). Second language speech learning theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–277). York Press.

Flege, J. (1999). Age of learning and second-language speech. In D. Birdsong (Ed.), *Second language acquisition and the critical period hypothesis* (pp. 101–132). Lawrence Erlbaum.

Flege, J., & Bohn, O.-S. (2021). The revised speech learning model (SLM-r). In R. P. Wayland (Ed.), *Second language speech learning: Theoretical and empirical progress* (pp. 3–83). Cambridge University Press. https://doi.org/10.1017/9781108886901.002

Flege, J., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, *25*(4), 437–470. https://doi.org/10.1006/jpho.1997.0052

Flege, J., & Eefting, W. (1987a). Cross-language switching in stop consonant perception and production by Dutch speakers of English. *Speech Communication*, *6*(3), 185–202. https://doi.org/10.1016/0167-6393(87)90025-2

Flege, J., & Eefting, W. (1987b). Production and perception of English stops by native Spanish speakers. *Journal of Phonetics*, *15*(1), 67–83. https://doi.org/10.1016/S0095-4470(19)30538-8

Flege, J., MacKay, I. R., & Meador, D. (1999). Native Italian speakers' perception and production of English vowels. *The Journal of the Acoustical Society of America*, *106*(5), 2973–2987. https://doi.org/10.1121/1.428116

Flege, J., & Schmidt, A. M. (1995). Native speakers of Spanish show rate-dependent processing of English stop consonants. *Phonetica*, *52*(2), 90–111. https://doi.org/10.1159/000262062

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1338–1349. https://doi.org/10.1098/rstb.2011.0417

Forster, K., & Forster, J. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods Instruments and Computers*, *35*(1), 116–124. https://doi.org/10.3758/BF03195503

Franken, M. K., Acheson, D. J., McQueen, J. M., Eisner, F., & Hagoort, P. (2017). Individual variability as a window on production-perception interactions in speech motor control. *The Journal of the Acoustical Society of America*, *142*(4), 2007–2018. https://doi.org/10.1121/1.5006899

Georgeton, L., Paillereau, N., Landron, S., Jiayin, G., & Kamiyama, T. (2012). Analyse formantique des voyelles orales du français en contexte isolé: à la recherche d'une référence pour les apprenants de FLE [Formant analysis of French oral vowels in isolated contexts: In search of a reference for learners of FFL]. In L.

Besacier, B. Lacouteux, & G. Sérasset (Eds.), *Actes de la conférence conjointe JEP-TALN-RECITAL 2012* [Proceedings of the 2012 joint JEP-TALIN-RECITAL conference] (Vol. 1, pp. 145–152). http://www.aclweb.org/anthology/U/U12/F12-1019.pdf

Golestani, N., & Zatorre, R. J. (2009). Individual differences in the acquisition of second language phonology. *Brain and Language*, *109*(2–3), 55–67. https://doi.org/10.1016/j.bandl.2008.01.005

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders*, *39*(5), 350–365. https://doi.org/10.1016/j.jcomdis.2006.06.013

Hanulíková, A., Dediu, D., Fang, Z., Bašnaková, J., & Huettig, F. (2012). Individual differences in the acquisition of a complex L2 phonology: A training study. *Language Learning*, *62*(s2), 79–109. https://doi.org/10.1111/j.1467-9922.2012.00707.x

Hao, Y.-C., & de Jong, K. (2016). Imitation of second language sounds in relation to L2 perception and production. *Journal of Phonetics*, *54*, 151–168. https://doi.org/10.1016/j.wocn.2015.10.003

Hattori, K., & Iverson, P. (2010, September 22–24). Examination of the relationship between L2 perception and production: An investigation of English/r/-/l/perception and production by adult Japanese speakers. In M. Nakano (Chair), *Second language studies: Acquisition, learning, education and technology* [Workshop]. Interspeech 2010, Tokyo, Japan. https://www.isca-speech.org/archive_v0/L2WS_2010/papers/lw10_P2-4.pdf

Henderson, A., Frost, D., Tergujeff, E., Kautzsch, A., Murphy, D., Kirkova-Naskova, A., Waniek-Klimczak, E., Levey, D., Cunnigham, U., & Curnick, L. (2012). The English pronunciation teaching in Europe survey: Selected results. *Research in Language*, *10*(1), 5–27. https://doi.org/10.2478/v10015-011-0047-4

Inceoglu, S. (2019). Individual differences in L2 speech perception: The role of phonological memory and lipreading ability. *The Modern Language Journal*, *103*(4), 782–799. https://doi.org/10.1111/modl.12591

Jachs, B., Blanco, M. J., Grantham-Hill, S., & Soto, D. (2015). On the independence of visual awareness and metacognition: A signal detection theoretic analysis. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(2), 269–276. https://doi.org/10.1037/xhp0000026

JASP team. (2020). *JASP* (Version 0.14) [Computer software]. https://jasp-stats.org

Jia, G., Strange, W., Wu, Y., Collado, J., & Guan, Q. (2006). Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2 exposure. *The Journal of the Acoustical Society of America*, *119*(2), 1118–1130. https://doi.org/10.1121/1.2151806

Kang, O., Rubin, D., & Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Language Testing*, *36*(4), 481–504. https://doi.org/10.1177/0265532219849522

Kartushina, N., & Frauenfelder, U. H. (2014). On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation. *Frontiers in Psychology*, *5*, Article 1246. https://doi.org/10.3389/fpsyg.2014.01246

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2015). The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. *The Journal of the Acoustical Society of America*, *138*(2), 817–832. https://doi.org/10.1121/1.4926561

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2016). Mutual influences between native and non-native vowels in production: Evidence from short-term visual articulatory feedback training. *Journal of Phonetics*, *57*, 21–39. https://doi.org/10.1016/j.wocn.2016.05.001

Kartushina, N., & Martin, C. D. (2019). Talker and acoustic variability in learning to produce nonnative sounds: Evidence from articulatory training. *Language Learning*, *69*(1), 71–105. https://doi.org/10.1111/lang.12315

Kartushina, N., Soto, D., & Martin, C. (2022a). *Analysis code. Datasets from "Metacognition in second language speech perception and production"* [Analysis code]. IRIS Database, University of York, UK. https://doi.org/10.48316/7jrb-g995

Kartushina, N., Soto, D., & Martin, C. (2022b). *Data. Datasets from "Metacognition in second language speech perception and production*" [Dataset]. IRIS Database, University of York, UK. https://doi.org/10.48316/jawv-s681

Kartushina, N., Soto, D., & Martin, C. (2022c). *Instructions and audio stimuli. Materials from "Metacognition in second language speech perception and production"* [Collection: stimuli]. IRIS Database, University of York, UK. https://doi.org/10.48316/asy8-wg12

Kluge, D. C., Rauber, A. S., Reis, M. S., & Bion, R. A. H. (2007). The relationship between the perception and production of English nasal codas by Brazilian learners of English. *Proceedings of Interspeech 2007*, 2297–2300. https://doi.org/10.21437/Interspeech.2007-623

Koriat, A. (2019). Confidence judgments: The monitoring of object-level and same-level performance. *Metacognition and Learning*, *14*(3), 463–478. https://doi.org/10.1007/s11409-019-09195-7

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General*, *135*(1), 36–69. https://doi.org/10.1037/0096-3445.135.1.36

Koriat, A., & Shitzer-Reichert, R. (2002). Metacognitive judgments and their accuracy. In P. Chambres, M. Izaute, & P.-J. Marescaux (Eds.), *Metacognition: Process, function and use* (pp. 1–17). Springer. https://doi.org/10.1007/978-1-4615-1099-4_1

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

Kuhlmann, B. G. (2019). Metacognition of prospective memory: Will I remember to remember? In J. Rummel & M. A. McDaniel (Eds.), *Prospective memory* (pp. 1–18). Routledge. https://doi.org/10.4324/9781351000154-5

Levy, E. S. (2009). Language experience and consonantal context effects on perceptual assimilation of French vowels by American-English learners of French. *The Journal of the Acoustical Society of America*, *125*(2), Article 1138. https://doi.org/10.1121/1.3050256

Levy, E. S., & Law, F. F. (2010). Production of French vowels by American-English learners of French: Language experience, consonantal context, and the perception-production relationship. *The Journal of the Acoustical Society of America*, *128*(3), Article 1290. https://doi.org/10.1121/1.3466879

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

Li, J. J., Ayala, S., Harel, D., Shiller, D. M., & McAllister, T. (2019). Individual predictors of response to biofeedback training for second-language production. *The Journal of the Acoustical Society of America*, *146*(6), 4625–4643. https://doi.org/10.1121/1.5139423

Llompart, M., & Reinisch, E. (2018). Acoustic cues, not phonological features, drive vowel perception: Evidence from height, position and tenseness contrasts in German vowels. *Journal of Phonetics*, *67*, 34–48. https://doi.org/10.1016/j.wocn.2017.12.001

MacIntyre, P. D., Noels, K. A., & Clément, R. (1997). Biases in self-ratings of second language proficiency: The role of language anxiety. *Language Learning*, *47*(2), 265–287. https://doi.org/10.1111/0023-8333.81997008

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, *2*(1), 49–55.

Mairano, P., Bouzon, C., Capliez, M., & De Iacovo, V. (2019). Acoustic distances, Pillai scores and LDA classification scores as metrics of L2 comprehensibility and nativelikeness. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.) *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 1104–1108). Australasian Speech Science and Technology Association, Inc. https://assta.org/proceedings/ICPhS2019/papers/ICPhS_1153.pdf

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430. https://doi.org/10.1016/j.concog.2011.09.021

McCurdy, L. Y., Maniscalco, B., Metcalfe, J., Liu, K. Y., Lange, F. P. de, & Lau, H. (2013). Anatomical coupling between distinct metacognitive systems for memory and visual perception. *Journal of Neuroscience*, *33*(5), 1897–1906. https://doi.org/10.1523/JNEUROSCI.1890-12.2013

Melnik-Leroy, G. A., Turnbull, R., & Peperkamp, S. (2021). On the relationship between perception and production of L2 sounds: Evidence from Anglophones' processing of the French /u/–/y/contrast. *Second Language Research*, *8*(3), 581–605. https://doi.org/10.1177/0267658320988061

Mitterer, H., Eger, N. A., & Reinisch, E. (2020). My English sounds better than yours: Second-language learners perceive their own accent as better than that of their peers. *PLOS ONE*, *15*(2), Article e0227643. https://doi.org/10.1371/journal.pone.0227643

Morales, J., Lau, H., & Fleming, S. M. (2018). Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *The Journal of Neuroscience*, *38*(14), 3534–3546. https://doi.org/10.1523/JNEUROSCI.2360-17.2018

Nagle, C. L. (2018). Examining the temporal structure of the perception–production link in second language acquisition: A longitudinal study. *Language Learning*, *68*(1), 234–270. https://doi.org/10.1111/lang.12275

Nagle, C. L., & Baese-Berk, M. M. (2021). Advancing the state of the art in L2 speech perception-production research: Revisiting theoretical assumptions and methodological practices. *Studies in Second Language Acquisition*, *44*(2), 580–605. https://doi.org/10.1017/S0272263121000371

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–25). The MIT Press. https://doi.org/10.7551/mitpress/4561.001.0001

Okuno, T., & Hardison, D. M. (2016). Perception-production link in L2 Japanese vowel duration: Training with technology. *Language Learning & Technology*, *30*(2), 61–80. https://doi.org/10125/44461

Peperkamp, S., & Bouchon, C. (2011). The relation between perception and production in L2 phonological processing. *Proceedings of Interspeech 2011*, 161–164. https://doi.org/10.21437/Interspeech.2011-72

Peters, M. A. K., & Lau, H. (2015). Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. *ELife*, *4*, Article e09651. https://doi.org/10.7554/eLife.09651

Quintana, D. S., & Williams, D. R. (2018). Bayesian alternatives for common null-hypothesis significance tests in psychiatry: A non-technical guide using JASP. *BMC Psychiatry*, *18*(1), Article 178. https://doi.org/10.1186/s12888-018-1761-4

Rahnev, D., Desender, K., Lee, A. L. F., Adler, W. T., Aguilar-Lleyda, D., Akdoğan, B., Arbuzova, P., Atlas, L. Y., Balcı, F., Bang, J. W., Bègue, I., Birney, D. P., Brady, T. F., Calder-Travis, J., Chetverikov, A., Clark, T. K., Davranche, K., Denison, R. N., Dildine, T. C., … Zylberberg, A. (2020). The confidence database. *Nature Human Behaviour*, *4*(3), 317–325. https://doi.org/10.1038/s41562-019-0813-1

Rallo Fabra, L., & Romero, J. (2012). Native Catalan learners' perception and production of English vowels. *Journal of Phonetics*, *40*(3), 491–508. https://doi.org/10.1016/j.wocn.2012.01.001

Rouault, M., McWilliams, A., Allen, M., & Fleming, S. (2018). Human metacognition across domains: Insights from individual differences and neuroimaging. *Personality Neuroscience*, *1*, Article E17. https://doi.org/10.1017/pen.2018.16

Reid, K. T., Trofimovich, P., & O'Brien, M. G. (2019). Social attitudes and speech ratings: Effects of positive and negative bias on multiage listeners' judgments of second language speech. *Studies in Second Language Acquisition*, *41*(2), 419–442. https://doi.org/10.1017/S0272263118000244

Saito, K., Tran, M., Suzukida, Y., Sun, H., Magne, V., & Ilkan, M. (2019). How do second language listeners perceive the comprehensibility of foreign-accented speech? Role of first language profiles, second language proficiency, age, experience, familiarity and metacognition. *Studies in Second Language Acquisition*, *41*(5), 1133–1149. https://doi.org/10.1017/S0272263119000226

Samaha, J., & Postle, B. R. (2017). Correlated individual differences suggest a common mechanism underlying metacognition in visual perception and visual short-term memory. *Proceedings of the Royal Society B: Biological Sciences*, *284*(1867), Article 20172035. https://doi.org/10.1098/rspb.2017.2035

Sato, M., & Loewen, S. (2018). Metacognitive instruction enhances the effectiveness of corrective feedback: Variable effects of feedback types and linguistic targets. *Language Learning*, *68*(2), 507–545. https://doi.org/10.1111/lang.12283

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of Phonetics*, *52*, 183–204. https://doi.org/10.1016/j.wocn.2015.07.003

Sheldon, A., & Strange, W. (1982). The acquisition of /r/and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics*, *3*(3), 243–261. https://doi.org/10.1017/S0142716400001417

Simmonds, A. J. (2015). A hypothesis on improving foreign accents by optimizing variability in vocal learning brain circuits. *Frontiers in Human Neuroscience*, *9*, Article 606. https://doi.org/10.3389/fnhum.2015.00606

Simmonds, A. J., Wise, R. J. S., Dhanjal, N. S., & Leech, R. (2011). A comparison of sensory-motor activity during speech in first and second languages. *Journal of Neurophysiology*,*106*(1), 470–478. https://doi.org/10.1152/jn.00343.2011

Song, J. Y., & Eckman, F. R. (2021). The relationship between second-language learners' production and perception of English vowels: The role of native-like acoustic correlates. In D. Dionne & L.-A. Vidal Covas (Eds.), *Proceedings of the 45th Annual Boston University Conference on Language Development* (pp. 693–706). Cascadilla Press.

Strachan, L., Kennedy, S., & Trofimovich, P. (2019). Second language speakers' awareness of their own comprehensibility: Examining task repetition and self-assessment. *Journal of Second Language Pronunciation*, *5*(3), 347–373. https://doi.org/10.1075/jslp.18008.str

Traunmüller, H. (1997). *Auditory scales of frequency representation* [Online tutorial]. http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-10230

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

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America*, *113*(2), 1033–1043. https://doi.org/10.1121/1.1531176

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*(2), 231–252. https://doi.org/10.1177/0265532212456968

Yule, G., Damico, J., & Hoffman, P. (1987). Learners in transition: Evidence from the interaction of accuracy and self-monitoring skill in a listening task. *Language Learning*, *37*(4), 511–521. https://doi.org/10.1111/j.1467-1770.1987.tb00582.x

Yule, G., Hoffman, P., & Damico, J. (1987). Paying attention to pronunciation: The role of self-monitoring in perception. *TESOL Quarterly*, *21*(4), 765–768. https://doi.org/10.2307/3586994

Zhang, K., & Peng, G. (2017). The relationship between the perception and production of non-native tones. *Proceedings of Interspeech 2017*, 1799–1803. https://doi.org/10.21437/Interspeech.2017-714

Supporting Information

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

Accessible Summary

Appendix S1. Settings for Linear Regression and Logistic Regression.

Appendix S2. Intersubject Correlations Between the Mean Confidence Ratings in the Perception and Production Tasks.