Studies in Second Language Acquisition, **44** (2022), 580–605
doi:10.1017/S0272263121000371

State of the Scholarship

ADVANCING THE STATE OF THE ART IN L2 SPEECH PERCEPTION-PRODUCTION RESEARCH: REVISITING THEORETICAL ASSUMPTIONS AND METHODOLOGICAL PRACTICES

Charles L. Nagle *

Iowa State University

Melissa M. Baese-Berk

University of Oregon

Abstract

One of the basic goals of second language (L2) speech research is to understand the perception-production link, or the relationship between L2 speech perception and L2 speech production. Although many studies have examined the link, they have done so with strikingly different conceptual foci and methods. Even studies that appear to use similar perception and production tasks often present nontrivial differences in task characteristics and implementation. This conceptual and methodological variation makes meaningful synthesis of perception-production findings difficult, and it also complicates the process of developing new perception-production models that specifically address how the link changes throughout L2 learning. In this study, we scrutinize theoretical and methodological issues in perception-production research and offer recommendations for advancing theory and practice in this domain. We focus on L2 sound learning because most work in the area has focused on segmental contrasts.

INTRODUCTION

Cognitive scientists have long been interested in the relationship between perception and action, or perception-action links. One such link that has been a topic of considerable focus in speech research, especially second language (L2) speech research, is the

*Correspondence concerning this article should be addressed to Charles L. Nagle, Iowa State University, Department of World Languages and Cultures, 3102 Pearson Hall, 505 Morrill Road, Ames, Iowa 50011. E-mail: cnagle@iastate.edu

relationship between speech perception and speech production. Although is it uncontroversial that the two modalities are related, researchers have approached the perception-production link from diverse theoretical perspectives. Some have argued that speech perception should be conceptualized as a domain-general categorization problem (Chandrasekaran et al., 2014), while others have advocated for domain-specific approaches. Within L2 sound learning, several models, including the Perceptual Assimilation Model (PAM; Best, 1995) and its L2 extension (PAM-L2, Best & Tyler, 2007) and the Speech Learning Model (SLM; Flege, 1995) and its recent revision (SLM-r; Flege & Bohn, 2021) have been developed to describe how the phonetic and phonological organization of the native language (L1) influences the perception and production of L2 sounds. Accordingly, research conducted within these frameworks has yielded rich insight into how L1 and L2 sound systems interact in various learner populations and at various points in L2 learning.

At the same time, these models have been widely applied as general perception-production frameworks, even though they were not designed to explain how perception-production links develop and evolve throughout the L2 learning process. The overgeneralization of these models is problematic for several reasons. First, in the absence of specific and testable perception-production hypotheses, research in this area has generally evaluated the broad (and uncontroversial) claim that accuracy in perception is related to accuracy in production, a research agenda that is unlikely to advance current theory and practice. Second, because models are understandably silent with respect to perception-production methodology (but cf. Flege & Bohn, 2021), researchers have used a wide variety of perception and production tests to investigate this relationship. As a result, as a field, we have accumulated a large body of perception-production research, but studies may not be methodologically robust or valid (e.g., because they use tests that do not tap into the underlying perception and production skills they are designed to measure or analyze perception and production measures that should not be equated), and even if they are, they may prove orthogonal to one another (e.g., because they examine distinct facets of the perception-production relationship or distinct learner populations).

The time is right to take a step back and appraise current approaches to the perception-production link. For one, the fact that current models have undergone revision highlights that work in this area has and should continue to evolve, making room for new theoretical insights and methodological approaches. What's more, advances in data collection, management, and processing, as well as a renewed emphasis on advanced statistical methods in L2 research (e.g., Plonsky, 2015) has catalyzed methodological innovation in all areas of second language acquisition research, and the same is true of research on the perception-production link. There are also practical reasons to scrutinize perception-production work. Notably, improving the quality of perception-production research and expanding its scope can suggest new ways to optimize speech training (see, e.g., Sakai & Moorman, 2018). With this in mind, in this state of the scholarship paper, we briefly review existing models of L2 sound learning before turning to theoretical and methodological issues in examining the perception-production link. We then conclude with recommendations for conducting research in this domain. We focus on segmental perception-production research because most of the work that has informed current models is segmental in nature.[1]

https://doi.org/10.1017/S0272263121000371 Published online by Cambridge University Press

BACKGROUND

MODELS OF L2 SPEECH PERCEPTION AND PRODUCTION

Researchers have proposed a variety of models to account for L2 speech learning, but most of these models were designed to account for performance and/or learning in only one modality. For instance, PAM (Best, 1995) was initially developed to explain how naïve listeners assimilate nonnative sounds to native language categories. Best and Tyler (2007) later applied this model to L2 learning, arguing that perceptual assimilation patterns determine the degree of difficulty L2 listeners experience in learning to perceive L2 contrasts. According to PAM-L2, single category assimilations, where both members of the L2 contrast are assimilated to a single L1 category and perceived to be relatively equal in terms of their goodness of fit, should be especially challenging for L2 learners. On the other hand, two category assimilations, where contrastive L2 sounds are assimilated to distinct L1 categories, should be comparatively easy to learn. PAM and PAM-L2 are built on a direct realist approach to speech perception (e.g., Fowler, 1986), which assumes that listeners directly perceive articulatory gestures. Thus, within this framework, perception and production should be closely intertwined even though neither PAM nor PAM-L2 directly addresses the perception-production link. What's more, PAM-L2 lacks a strong developmental component, making it difficult to derive testable, longitudinal claims about perception-production relationships.

Another model of L2 perception is the L2 Linguistic Perception model (L2LP; van Leussen & Escudero, 2015), which is grounded in an optimality-based approach to perceptual learning. According to this model, learning is error-driven. When a learner becomes aware of a perceptual error, pathways between phonetic, phonological, and lexical tiers of representation are altered to reduce the likelihood that a similar error will occur in the future. The L2LP is intended to account for the entirety of L2 learning. Yet, like PAM, it does not directly address L2 speech production, nor does it include information on the perception-production link, although presumably accurate perception would be a necessary precondition for accurate production (Escudero, 2007). Cognitive scientists working outside of mainstream SLA research have also developed models to explain perceptual learning. For example, Chandrasekaran et al. (2014) argued that L2 speech category learning should be viewed as a general categorization problem involving both reflective and reflexive (i.e., explicit and implicit) learning systems. Overall, then, these models are useful for understanding L2 perception, but their implications for L2 production and the perception-production link are fuzzy.

With respect to speech production, the SLM (Flege, 1995) has received far more attention than any other framework. This model was based on findings showing that even individuals who had immigrated to an L2 environment at a young age had a noticeable foreign accent in the L2 and produced L2 sounds whose phonetic characteristics did not align with those of age-matched monolinguals. The overarching aim of the SLM was, therefore, to explicate the relationship between age of onset of L2 learning, experiential variables such as quantity and quality of L2 input, and L2 pronunciation attainment. One of its key hypotheses was that many production errors have a perceptual basis. More specifically, according to this framework, individuals will be able to produce phonetically accurate L2 sounds only if they have formed a new phonetic category for those sounds. Although phonetic learning remains possible throughout the lifespan (which means that

https://doi.org/10.1017/S0272263121000371 Published online by Cambridge University Press

even late-start learners can modify their production), the likelihood of forming a new phonetic category is hypothesized to decrease as age of onset increases. This is because as the L1 becomes more robust, learners will have more and more difficulty discerning the subtle, yet important, phonetic differences that exist between crosslinguistically similar sounds. If learners detect these differences, then they may form a new phonetic category for the L2, which should enable, but not necessarily guarantee, accurate L2 production. If, on the other hand, they do not detect differences between L1 and L2 sounds, then they will associate the L2 sound with the L1 category, leading to accented L2 productions.

The original formulation of the SLM was squarely focused on phonetic ultimate attainment in highly proficient L2 users and posited a unidirectional pathway of accuracy in perception shaping accuracy in production. In contrast, in the revised model (SLM-r), Flege and Bohn (2021) have posited a bidirectional, co-evolving perception-production link, which means that perception and production should mirror one another (i.e., should be somewhat synchronized) during L2 learning. The SLM-r also moves away from an emphasis on phonological end states, favoring instead a developmental approach that would entail tracking when learners begin to discern differences between phonetically similar L1 and L2 sounds and how doing so facilitates the formation of new L2 phonetic categories.

However, as Flege and Bohn have acknowledged, the time course of L2 category formation is not well understood, nor are the events that potentially catalyze it: "At a later and as-yet undefined moment in phonetic development, the perceptual link between the L2 'equivalence' class and the L1 category will be sundered. We speculate that this delinking may be speeded by growth of the L2 lexicon, at least in literate learners of an L2" (2021, p. 26). The SLM-r also addresses a range of learner differences (e.g., auditory processing) that could account for variation in development.

Although the SLM(-r) offers the most robust starting point for evaluating perception-production links in L2 learning, there are many critical questions that the model does not fully address. For one, the notion of crosslinguistically similar sounds encompasses a range of qualitatively distinct learning targets and crosslinguistic relationships. In some cases, both the L1 and the L2 might contain the same phonological categories, and those categories might be implemented using the same phonetic cue, but precise perceptual crossover boundaries and production values differ. In other cases, the L2 might contain a three-way phonological contrast where the L1 has only a two-way contrast, and the L2 contrast may be implemented using different phonetic cues. Minimally, it would be reasonable to expect different developmental timelines for each scenario. It could also be the case that the functional form of the perception-production relationship varies according to such crosslinguistic relationships.

Finally, it remains an open question whether individual differences affect the speed with which perception and production become aligned and the strength of that alignment. Overall, then, to enhance current models and develop a comprehensive understanding of perception-production relationships in L2 learning, more research is warranted in three areas: tracking the link over time, examining the link for different types of crosslinguistically similar sounds, and investigating individual differences in each modality and in the link itself. In the following sections, we outline how research in each of these areas can inform theory.

https://doi.org/10.1017/S0272263121000371 Published online by Cambridge University Press

THEORETICAL ISSUES IN PERCEPTION-PRODUCTION RESEARCH

DEVELOPMENTAL STAGE

One issue that should be at the forefront of perception-production research is L2 speakers' developmental stage. Developmental stage is important because the perception-production relationship is time-varying. Conceptually, this means that a complete understanding of the link rests upon examining how it changes over time, that is, how perception and production interact as L2 speakers are exposed to different types of input and engage in varying levels of L1 and L2 use in different learning environments. On a practical level, an inherently time-varying link means that cross-sectional studies may over- or underestimate the link depending on the precise moment at which L2 speakers are measured, leading to a narrow or truncated view of perception-production relationships. Put another way, interpreting cross-sectional perception-production findings as representative of the underlying nature of the perception-production relationship risks reducing a dynamic developmental phenomenon to a static snapshot.

There are several studies that suggest that the perception-production link changes over time. For instance, Rallo Fabra, and Romero (2012) examined L1 Catalan speakers' perception and production of English vowels. They reported an overall perception-production correlation of $r = .26$, a small effect that was not statistically significant. However, when they evaluated the link within three distinct learner proficiency groups, they found a large perception-production correlation ($r = .76$) in intermediate L2 speakers. In another study on L1 Mandarin speakers' perception and production of English vowels, Jia et al. (2006) compared three groups: foreign language learners in China, second language learners who had lived in the US for less than 2 years, and another group of second language learners who had lived in the US for 3–5 years. The two second language groups were matched across a range of demographic variables, including age of arrival and age of onset of L2 instruction. The overall perception-production correlation was $r = .50$, and correlations for the foreign language learners and past arrivals were of similar magnitude ($r = .42$ and $r = .46$, respectively). However, for recent arrivals, the perception-production link was far weaker ($r = .25$). Although the goal of the study was to examine age and experience-related changes in the perception and production of L2 sounds, the between-group differences in the strength of the perception-production link are intriguing, insofar as they suggest distinct degrees of perception-production synchronization. In fact, the absence of a strong correlation in the recent arrival group could be interpreted as evidence of a change in the link. For example, a weak correlation could signal a lagged relationship, in which case cross-lagged measures would show a stronger relationship than their time-locked counterparts. To that point, longitudinal developmental studies—studies that observed perception-production relationships over time without providing training in either modality—point to a lagged model (Casillas, 2020a, 2020b; Nagle, 2018).

These intriguing findings underscore the need for a time-sensitive view of the perception-production link that is informed by environmental (e.g., the type of input that learners are exposed to) and speaker (e.g., the frequency and quality of L2 interactions) variables. Cross-sectional studies can be valuable, serving as a first step toward identifying the variables that should be measured longitudinally, but they are limited in terms of the perception-production questions that can be examined. Cross-sectional research can

https://doi.org/10.1017/S0272263121000371 Published online by Cambridge University Press

speak to the strength of perception-production relationships in different learner populations, but it cannot shed light on the developmental questions that have the greatest potential to advance theory and practice: Precisely when does perception being to influence production? More specifically, is there a dynamic coupling between perception and production, with changes in perception mirrored in production, or must perception accuracy reach a certain threshold before production begins to improve? To what extent does rate of change in perception predict rate of change in production? And how does the strength of the perception-production relationship change over time?

From a dynamic perspective, it is easy to imagine how perception-production relationships might change. For example, at the outset of learning, perhaps perception and production improve relatively quickly, with most learners reaching a moderate level of accuracy in both modalities within the first year of intensive L2 exposure (a period commonly referred to as the window of maximal opportunity for pronunciation learning; see Derwing & Munro, 2015). During this period, a large correlation between (cross-lagged) perception and production measures might be observed, with the correlation increasing in strength over time. Perception accuracy might even be the strongest single predictor of production accuracy. Once perception begins to stabilize, entering a developmental plateau, the perception-production link itself might also begin to stabilize, such that no change is observed in the link during, or the effect of perception on production may decrease in strength or disappear altogether.

This stasis might continue until a new experience, such as targeted pronunciation training, catalyzes additional development in either modality. At that point, a similar developmental cycle might ensue: perception and production improve, albeit at different rates, resulting in a relatively strong cross-lagged perception-production link that decays over time as perception fades from the strongest predictor, to one of many predictors, to having little predictive value at all. This scenario is of course hypothetical, but the point is that a developmental approach necessitates a consideration of the timing, strength, and duration of the effect of perception on production. It also demands careful consideration of what time-varying predictors should be sampled. For example, if perception and production learning are lexically driven, as many scholars have suggested (Best & Tyler, 2007; Flege, 1995; Flege & Bohn, 2021), then examining changes in vocabulary size over time would be important.

Thus far, we have focused on a macro-level perspective of the perception-production link, discussing how perception and production change and influence one another on a timescale of months or years. However, perception and production might also display substantial within-subjects variation on much shorter timescales, making it possible to distinguish between broad developmental processes and state-like variation in the perception-production link. A state-level view of the link opens up a new domain of research questions related to the stability of the perception and production systems. For instance, do individual differences in L1 and L2 use trigger temporary variation in perception and production accuracy? Although it can be tempting to attribute variation in either modality to measurement error, variation may in fact be indicative of the volatility of the developing systems. Even in advanced L2 users, it seems reasonable to expect perception, production, and the perception-production link to vary in response to varying patterns of L1 and L2 use. Perhaps L1 use temporarily disrupts, or even decouples, L2 perception and production, in the same way that producing sounds during

https://doi.org/10.1017/S0272263121000371 Published online by Cambridge University Press

perceptual training can disrupt perceptual learning (Baese-Berk, 2019; Baese-Berk & Samuel, 2016).

In summary, the current state of perception-production research has been informed by a large number of cross-sectional studies, which are not well-suited to capture the time-varying nature of the perception-production link (Nagle, 2021). What is needed then, is a departure from current methods and a reorientation toward a more dynamic developmental approach that is predicated on longitudinal sampling. Such sampling strategies hold the key to understanding both how perception and production change over time and how changes in perception relate to changes in production after controlling for other important developmental phenomena (e.g., vocabulary size, patterns of L1 and L2 use).

LEARNING SCENARIO

A deeper understanding of how the nature of the learning scenario/target structure affects development in each modality and the perception-production link can enhance models of L2 sound learning and inform the development of optimal instructional practices. Current models assume that L1 sound patterns determine how learners perceive and produce L2 sounds. Crosslinguistically similar L2 sounds are viewed as especially challenging because L2 learners may associate those sounds with L1 categories. Yet, crosslinguistic phonetic similarity as a theoretical construct encompasses a variety of qualitatively distinct subcategories, each of which could be characterized by different perception-production relationships. Practically speaking, this means that comparing studies that examine different subclasses of crosslinguistic similarity would be akin to comparing apples and oranges, which could explain divergent results in the literature. Put another way, learning scenario is likely to be an important moderator variable that must be accounted for in perception-production research.

One common scenario a learner is faced with is what the L2LP calls a boundary shift (van Leussen & Escudero, 2015). In this scenario, both the L1 and the L2 contain the same number of phonological categories, and those categories are phonetically implemented using the same cue, but the phonetic boundary is not the same in the two languages. For instance, English and Spanish both contain a two-way stop consonant voicing contrast (e.g., /b/-/p/) that is predominantly cued by differences in voice onset time (Lisker & Abramson, 1964). However, they differ with respect to the crossover boundary in perception and average voice onset time values in production. In English, phonologically voiced stops are phonetically realized as prevoiced or short-lag, whereas phonologically voiceless stops are realized as long-lag (/b/ is implemented as either [b] or [p], whereas /p/ is implemented as [p^h], except after /s/). This contrasts with Spanish, where voiced stops are prevoiced and voiceless stops short-lag (/b/ is implemented as [b] and /p/ as [p]). Thus, an English speaker unfamiliar with Spanish would likely perceive Spanish voiced and voiceless stops as instances of English voiced stops. To perceive Spanish stops correctly —and produce them correctly, assuming that accurate perception is one of the primary determinants of accurate production—English speakers would need to learn to distinguish between prevoiced and short-lag variants, associating prevoiced variants with phonological voicing and short-lag variants with phonological voicelessness. As this example makes apparent, this might involve retuning the relationship between phonetics and

https://doi.org/10.1017/S0272263121000371 Published online by Cambridge University Press

phonology to match the L2, or from an SLM(-r) perspective, developing new L2 phonetic categories for Spanish stops.

A second learning scenario is when both languages use the same phonetic cue but the L2 contains more phonological categories than the L1. In this case, learning entails establishing an entirely new phonological category in a region of phonetic space where there is only one category in the L1. For example, Thai has a three-way stop consonant voicing contrast that, like the two-way stop contrast in English, is cued by differences in voice onset time (i.e., same phonetic cue, different number of phonological categories). In this case, English speakers would need to create separate categories for prevoiced and short-lag stops, both of which correspond to English voiced stops (i.e., they need to create a three-way /b/-/p/-/p^h/ distinction). Creating a new category is qualitatively distinct from shifting boundaries or retuning the phonetics-phonology interface to match the L2. Thus, different perception-production relationships could arise for L2 Spanish and L2 Thai stops.

Learners may sometimes be faced with a scenario in which the L2 contains more phonological categories than the L1 and implements those categories using a novel cue. This would be the case for English and Korean stops, given that Korean contains a three-way stop consonant contrast that is jointly cued by voice onset time and fundamental frequency in the following vowel (see, e.g., Schertz et al., 2015). As a result, English speakers would need to learn to attend to a novel phonetic cue that is not primary in the L1 to create accurate perceptual representations for Korean stop consonants. Whereas English speakers learning Thai might be able to begin developing new categories relatively quickly because the two languages make use of the same phonetic cue, it is reasonable to hypothesize that perceptual learning in Korean would proceed more gradually given the presence of a novel cue. These crosslinguistic relationships are exemplified in Figure 1.

It is important to note that the examples above are only one set of possibilities for how the phonetic space may be different between the L1 and L2. All of these cases reflect situations where multiple categories exist in each language. It is also often the case that the L1 contains only one category where the L2 has two. Two commonly researched

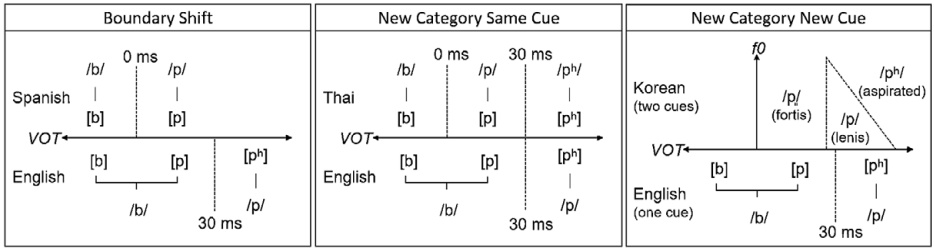

FIGURE 1. Crosslinguistic comparison of learning scenarios: English vs. Spanish, Thai, and Korean.
Note. For English, Spanish, and Thai, the primary phonetic cue to the stop consonant contrast is voice onset time. For Korean, the stop consonant contrast is cued by both voice onset time and fundamental frequency (f_0); hence, the two-dimensional space for the new category new cue example involving Korean (f_0 is shown on the vertical axis, and voice onset time on the horizontal axis). Phonological categories appear between forward slashes (/b/) and phonetic categories between brackets ([b]).

https://doi.org/10.1017/S0272263121000371 Published online by Cambridge University Press

examples are Japanese speakers' perception and production of the English /r/-/l/ contrast and non-native English speakers' perception and production of English tense-lax vowel contrasts such as /i/-/ɪ/. Both of these examples would require learners to begin attending to the relevant phonetic cue in the L2. Although at face value they seem to line up well with the English/Korean stop consonant example (i.e., adding a new category using a new cue), in fact, there is a critical difference: in one-to-two category scenarios, there is not necessarily a preexisting cue in the L1 that could influence L2 perception, whereas in two-to-three category scenarios, there is a preexisting L1 cue that learners may need to retune to fit L2 characteristics. For instance, even though both Korean and English make use of voice onset time, the way the cue is used in the two languages is different. This means that English speakers would need to adjust their perception of voice onset time to fit Korean while simultaneously learning to attend to f_0, a novel cue.

Finally, there are also crosslinguistic differences in phonotactics. For example, English and Arabic show a similar voicing contrast for alveolar stops (/d/-/t/, phonetically [d]-[t^{h}]), but Arabic does not have the same contrast at the bilabial and velar places of articulation, where only /b/ and /k/ are instantiated at a phonological level. Thus, Arabic speakers of English would need to create a new phonological contrast, analogous to their native Arabic /t/-/d/ contrast, at two new places of articulation. Likewise, English contrasts initial, medial, and final stops, whereas many languages only show a stop consonant contrast word-initially or word-medially, as in Spanish. Moreover, each of these phonological environments is associated with a different set of cues: in English, voice onset time word-initially, voice onset time and closure duration word-medially, and duration of the preceding vowel word-finally.

In summary, as a conceptual category, crosslinguistic similarity includes a variety of meaningful subclasses, each of which could be associated with distinct perception-production patterns. Future perception-production work must systematically examine these subclasses, taking stock of crosslinguistic differences in phonetic, phonological, and phonotactic patterns, to examine how the perception-production link manifests across a range of learning scenarios.

INDIVIDUAL DIFFERENCES

Individual differences could impact the perception-production link in two different ways: individual differences in the nature of the link itself, in which case it would be appropriate to discuss qualitatively different perception-production links that arise due to individual differences in, for instance, auditory processing ability, or individual differences in the way the link (i.e., a single, invariant perception-production relationship) manifests. There is reason to believe that qualitatively distinct perception-production links might exist, perhaps even within the same learner at different points in time.

Chandrasekaran et al. (2014) have argued that L2 sound learning should be conceptualized as a general categorization problem that involves two neurobiologically distinct systems: a reflective system that relies on working memory to test explicit categorization rules and a reflexive system that associates perceptual stimuli with motor outputs. A dual-learning system approach has implications for the perception-production link. Namely, between- and within-subjects variation in reflective and reflexive strategy use could lead to different perception-production relationships. Some learners may show a greater

https://doi.org/10.1017/S0272263121000371 Published online by Cambridge University Press

reliance on one system than the other, and the same learner may show varying levels of reliance on the two systems at different points during L2 learning. A learner might initially use the reflective system before switching to the reflexive system. In fact, the learning environment may stimulate differences in system use. For instance, in many instructed contexts, learners receive explicit pronunciation instruction, which could encourage the use of reflective strategies.

Another possibility is that individual differences influence learning in each modality, which could in turn affect how the perception-production link manifests, that is, the rate at which perception and production become aligned and the strength and duration of that alignment. For example, Cerviño-Povedano and Mora (2010) found that learners with greater phonological short-term memory identified L2 vowels more accurately and were less reliant on secondary (i.e., less informative) phonetic cues in perception. In another study, Darcy et al. (2016) reported a significant link between inhibitory control and L2 vowel discrimination accuracy. These cross-sectional findings suggest that individual differences in cognitive skills might help learners develop accurate L2 phonological representations and block interference from L1 categories during L2 learning. However, this claim would need to be evaluated longitudinally to determine the extent to which these variables are associated with variation in perception-production learning over time.

Another candidate for research would be individual differences in language use. Although all models of L2 sound learning address this topic, L2LP (van Leussen & Escudero, 2015) and Simulation Theory (Gambi & Pickering, 2013) are particularly germane to the present discussion. If perceptual learning is error-driven, as van Leussen and Escudero (2015) suggested, and if perception facilitates production (Escudero, 2007), then learners who engage in the L2 frequently (quantity of L2 use) and extensively (quality of L2 use) might show more rapid development in each modality, and perhaps a stronger perception-production link, than learners whose patterns of L2 use are more circumscribed. L2 use is also central to Simulation Theory, according to which perception is grounded in the covert simulation of speech:

> First, a motor command is recovered using a combination of prior knowledge and perceptual input. This command constitutes the perceiver's representation of the goal underlying the observed unfolding action. Then, the perceiver derives the motor command that is most likely to follow, and feeds it into a forward model. The output of the forward model is the predicted sensory input if the motor command were executed. Predicted input can be compared to actual input (i.e., to a perception of the unfolding action) and the resulting "prediction error" can be used to adjust the motor command. (Gambi & Pickering, 2013, p. 4)

From this viewpoint, the more often an individual interacts in the L2, the more often they would have the opportunity to fine-tune perception-production relationships. A full presentation of Simulation Theory is beyond the scope of this study, but the point is that, according to this and other approaches, L2 use appears to be critical for establishing and adjusting associations between the acoustic signal and motor commands. Many studies have examined the relationship between L2 use and speech production accuracy, but to the best of our knowledge, no study has examined it in relation to the perception-production link itself.

Individual differences might also interact with pronunciation training paradigms, which could influence perception-production relationships through aptitude-treatment interactions. For example, Perrachione et al. (2011) found that high aptitude learners (individuals who performed well on pitch perception tests) performed well in high variability phonetic training, whereas for low aptitude learners, certain forms of high variability training were detrimental. Although not strictly an individual difference, one other mediating factor that deserves attention is the nature of the training paradigm itself. Baese-Berk and Samuel (2016) found that an integrated perception-production training paradigm had a negative impact on perceptual learning; perceptual gains were not as robust for the group of learners who had to produce sounds after hearing them compared with the group that engaged in perception-only training. One could imagine how individual differences (e.g., phonological short-term memory, inhibitory skill) might also play a role in mitigating the disruptive effect of production on perception in integrated training paradigms.

From a theoretical perspective, we have identified three key areas that future perception-production research should consider: developmental stage, learning scenario, and individual differences. In the next sections, we address methodological considerations, emphasizing the distinct methodologies that would be needed to carry out robust perception-production research.

METHODOLOGICAL ISSUES IN PERCEPTION-PRODUCTION RESEARCH

Developing valid, reliable, and comparable perception and production measures can prove difficult. Researchers must adopt a theoretical framework and decide on the underlying skills that they would like to measure in each modality. This means making choices about how accuracy in each modality should be defined and the level at which it should be measured. For instance, perception tasks routinely involve the presentation of syllables and/or individual words for discrimination, categorization, and identification, resulting in a relatively narrow view of perceptual skill as the ability to discriminate sounds under optimal conditions.

On the other hand, production tasks can involve reading words and phrases, repeating words and phrases, naming pictures, describing pictures, or responding to prompts, and accuracy can be defined at an acoustic level, in terms of phonetic features, or using listener-based judgments. Moreover, many perception measures represent sensitivity to contrast (i.e., multiple stimuli) and, therefore, encode an altogether different type of information than production measures, which focus on the ability to produce individual words or features accurately. Thus, perception-production research must contend with the very real possibility of comparing apples to oranges. If a longitudinal perspective is adopted, which in our view can afford the type of perception-production data that is most likely to enhance current models of L2 sound learning, researchers must also map tasks to anticipated developmental timelines. Finally, there is the issue of how to analyze the data. That is, researchers must decide not only on how to measure each skill, but also on what counts as robust evidence of a perception-production link, and if directionality is assumed (e.g., perception guides production), how to test for it.

MEASURING PERCEPTION

Speech perception is a complex cognitive process that is grounded in the integration of different types of information available at different levels of linguistic structure and memory (e.g., the speech signal itself, phonotactic probability, knowledge of the target variety or even the individual speaker). A variety of tasks have been used to measure speech perception, and each task may tap into slightly different aspects of a listener's ability to process new speech sounds. It is important to note that it is impossible to divorce the perceptual behaviors we are interested in measuring from the tasks we use to measure them. Therefore, in the current section, we explore methods that are often used to measure speech perception. For each measure, we address what participants are being asked to do, what the outcome measure of the task is, and what this outcome measure is likely to reflect.

We also discuss the stimuli that are used and how they are paired with tasks to yield categorical and gradient perspectives on perceptual processing. Each decision that is made in experimental design has an impact on what is being measured and how the resulting behavioral data can be interpreted. There is significant overlap in what tasks can tap into, so we believe it is important not to divide tasks into strictly independent bins and assume that all behavior within these bins reflects the same perceptual process.

Before we focus on perception tasks as a whole, it is important to consider stimuli characteristics because they can have an impact on the type of perceptual processing in which listeners engage. Stimuli can be naturally produced tokens or tokens drawn from a synthesized phonetic continuum. When exposed to naturally produced tokens that are canonical exemplars of target categories, listeners may respond in a more categorical way, drawing upon the rich acoustic variation present in the stimuli to make perceptual judgments. Likewise, when exposed to stimuli drawn from a continuum, they may respond in a finer-grained manner, directing their perceptual processing to precisely those dimensions along which stimuli vary.

Put another way, synthesized stimuli, which have been created by systematically varying target phonetic dimensions to create a series of steps, are tightly controlled, and, therefore, may elicit a more gradient response from the listener.[2] Similarly, stimuli drawn from multiple talkers may encourage more categorical performance than stimuli drawn from a single talker because of the abstraction processes required to interpret different acoustic signals as members of the same speech sound category. Even the inclusion of different types of filler trials can result in different perceptual performance. For example, Baese-Berk (2010) demonstrated that native English listeners appeared to be less sensitive to a novel contrast they had recently been trained on (i.e., prevoiced [da] vs. short-lag [ta]) when the test included filler trials corresponding to a known contrast (i.e., [ma] vs. [la]) than when the target stimuli were presented without filler trials. This suggests that what listeners interpret as phonologically and/or phonetically similar may depend on the types of contrasts to which they are exposed, including filler contrasts. Researchers should be sensitive to these issues (natural vs. synthetic, single vs. multiple talkers, filler items, etc.) when developing stimuli and, as laid out below, when pairing stimuli with different perceptual tasks.

Stimuli can be embedded in a variety of perceptual tasks, which means that stimuli and task characteristics necessarily interact to affect the type of processing in which listeners

https://doi.org/10.1017/S0272263121000371 Published online by Cambridge University Press

engage. Perceptual tasks can be roughly divided into two types: those that require listeners to compare multiple speech sounds and those that require listeners to engage with a single target sound at a time. Single-target tasks tend to be simpler, insofar as listeners hear a target stimulus (e.g., a sound, syllable, word) and are asked to make a judgment about it. This judgment may require the participant to match the input with an orthographic representation,[3] a picture, or a motor response (i.e., press right when you hear one sound, and left when you hear another). Regardless of the matching procedure, the fundamental task participants are being asked to complete is one of labelling or identification.

This type of task can be used to evaluate how listeners map auditory exemplars onto representations stored in long-term memory. As a result, it can provide information on the content of emerging perceptual categories. If the task is paired with naturally produced stimuli, it can provide insight into whether learners have begun to create separate perceptual categories for L2 sounds, and when paired with stimuli drawn from a phonetic continuum, it can shed light on phonetic cue use and cue weights. For instance, Schertz et al. (2015) created two sets of 141 stimuli varying along three acoustic dimensions to examine Korean speakers' perception (and production) of Korean and English stop consonants. Listeners heard each stimulus and were asked to classify it using a closed set of options. Similar tasks have been used to evaluate L2 listeners' reliance on spectral and duration cues in the perception of L2 vowel contrasts (e.g., Flege et al., 1997; Sakai, 2016) and to gain insight into the precise location of L2 phonemic boundaries, including how such boundaries change over time as a function of L2 experience (e.g., Casillas, 2020b).

Like their single-stimulus counterparts, perceptual tasks requiring listeners to compare multiple stimuli can take many forms. One common example is an AX discrimination task, where an anchor stimulus is presented (A) followed by a target stimulus (X), and the listener is instructed to decide if the two stimuli are the same or different. Another common task is an ABX categorization task. In this case, two anchor stimuli are presented (A and B) followed by a target stimulus (X), and the listener is asked to assign the target stimulus to one of the anchors. This task is more complex than AX discrimination because it entails goodness of fit comparisons (between X and A and X and B) and binning. That is, listeners need to group like sounds together, which suggests that they have at least tacit knowledge of the dimensions along which X and A/B differ.[4]

Oddity tasks are an amalgamation of both discrimination and categorization. On this task, three stimuli are presented, and the listener is asked to determine if they pertain to the same category or if there is an odd item out. If there is an odd item, they are asked to indicate its serial position (1, 2, or 3). As a result, on an oddity task, listeners must compare multiple pairs of stimuli (1 vs. 2, 2 vs. 3, and 1 vs. 3), deciding if they are the same or different, while simultaneously determining if the degree of correspondence between the pairs is roughly equal. These tasks provide insight into how sensitive listeners are to a given contrast. Yet, as with single-stimulus tasks, AX, ABX, and oddity tasks can be rendered more categorical or gradient depending on the stimuli that are used.

Furthermore, the length of the interstimulus interval has been shown to play a key role in the type of processing in which listeners engage (Schouten et al., 2003; Werker & Logan, 1985). Shorter intervals seem to encourage gradient (i.e., acoustic or phonetic) processing, whereas longer intervals are associated with categorical (i.e., phonological) processing. Because these tasks involve storing and comparing more than one stimulus,

individual differences in phonological short-term memory and auditory processing could also affect task performance. Importantly, the ability to discriminate similar sounds under certain conditions does not suggest the ability to identify those sounds correctly; discrimination and identification are not the same.

After developing perception tasks, researchers must decide how they will code and analyze the resulting data. Some tests are amenable to a single outcome measure (e.g., synthetic continua lend themselves to phonemic boundaries and cue weights), but others permit a range of options. Researchers are often interested in quantifying accuracy. Accuracy can be assessed at a trial level, as the probability of a correct response (e.g., Kartushina et al., 2015), or globally, as percent correct or a discrimination index. Sensitivity indices such as d-prime (*d'*) are advantageous because they take response bias into account.[5] Researchers might also choose to examine reaction times on correct response trials to gain insight into speed of phonological processing. It is important to determine which measure is most appropriate for the test and stimuli used, as these factors are quite likely to influence interpretation of the results. Of course, tasks, stimuli, and outcome measures should also be aligned with research questions.

To this point, we have demonstrated that the tasks used to measure perception vary substantially and that the outcomes being measured and interpreted may differ as a function of task specifics, including the stimuli used. We now shift our attention to how these tasks can be used to answer questions of development of L2 perception. Models of L2 sound learning suggest that multiple measures would be needed to understand L2 perceptual development. For instance, the SLM(-r) posits a specific sequence: learners need to become sensitive to fine-grained crosslinguistic differences before they can begin creating novel categories for L2 sounds. In that case, discrimination and categorization tests could be used to examine sensitivity to such differences, and identification could be used to evaluate the emergence of L2 categories.

PAM-L2 also assigns a central role to crosslinguistic similarity in shaping perceptual learning. According to this model, crosslinguistic perceptual assimilation patterns determine the relative difficulty of L2 contrasts. Research has shown that even learners from the same L1 background show variable patterns, which can affect the initial state of L2 perception and the difficulty that learners experience over time (Mayr & Escudero, 2010). Thus, a comprehensive investigative approach to L2 perceptual learning would include examining learners' L1-L2 perceptual assimilation patterns and tracking their discrimination, categorization, and identification performance longitudinally. Such an approach can provide insight into two key questions: Do individual differences in L1-L2 perceptual assimilation patterns affect the starting point for L2 perception? And does discrimination/categorization accuracy predict identification accuracy?

Indeed, a complete understanding of perceptual development during L2 learning rests upon examining how different perceptual processes unfold over time. Thus, defining perception based on learners' performance on a single perception test is theoretically and methodologically problematic. For one, different processes likely develop at different rates, which means that the right test must be used at the right time to accurately capture the state of learners' developing systems. Interpreting the results of any single perception test as representative of the entire system can lead to imprecise generalizations about the nature of perceptual learning. This imprecision is magnified when studies that use fundamentally different tests (e.g., discrimination vs. identification) or implement the

https://doi.org/10.1017/S0272263121000371 Published online by Cambridge University Press

same test in different ways are compared. In these cases, performance variation may be due to nontrivial differences in task characteristics. In summary, then, we propose that a variety of perception tasks must be coordinated and deployed to understand L2 perceptual learning. Notably, performance on discrimination and categorization tests should improve before performance on identification tests if discerning crosslinguistic sound differences is a necessary first step toward creating L2 categories. We also advocate for examining performance at various levels of granularity (e.g., by adopting both categorical and fine-grained perspectives on perceptual accuracy and learning).

MEASURING PRODUCTION

Like perception, production is a complex skill that can be measured and defined in a variety of ways. And like perception, different facets of production ability likely develop at different rates, which means that researchers must take care to select and sequence appropriate production tasks over appropriate developmental windows. In their measurement framework, Saito and Plonsky (2019) observed that production accuracy can be measured in controlled and spontaneous speech. They linked control production to declarative pronunciation knowledge and spontaneous production to procedural pronunciation knowledge. They also distinguished between three coding options that yield different perspectives on production accuracy: acoustic measurements, expert ratings of linguistic features, and listener intuition.

Although current models of L2 sound learning do not address how different facets of production ability develop over time, one can imagine a potential developmental sequence based on the factors that Saito and Plonsky identified. For instance, the ability to produce an intelligible L2 contrast might emerge first in controlled speech, when speakers can focus on their pronunciation. Then, they might begin producing an intelligible contrast in spontaneous speech (Saito, 2019). Over time, given enough input and the right combination of aptitude and interest, speakers might even begin to produce nativelike sounds, although most will fall short of that mark. In fact, even if L2 speakers are successful at mastering certain phonetic cues, they may struggle to achieve nativelike accuracy in all dimensions of L2 sound production, especially on spontaneous tasks (Saito, 2013) and when nativelike accuracy is scrutinized at a phonetic or acoustic level (Stölten et al., 2014).

To examine controlled production knowledge, researchers commonly rely on word- and sentence-level tasks, including word and sentence reading, word and sentence repetition, and picture naming. These tasks are not without their pitfalls. For instance, reading tasks do not exclusively test production ability, but rather literacy skills (i.e., phoneme-to-grapheme mappings). This is not trivial because many languages, including English, have an opaque orthographic system that learners may not have fully mastered, especially when pronouncing low frequency forms that do not conform to prototypical sound-spelling patterns. Repetition tasks avoid the confounding effect of orthography. At the same time, they confound perception and production skills.

Because listeners are exposed to an auditory stimulus that they must process before repeating it, inaccurate production may be a result of inaccurate perceptual processing rather than production difficulty. What's more, on an immediate repetition task, speakers may be able to store phonetic forms in short-term memory, potentially bypassing their

https://doi.org/10.1017/S0272263121000371 Published online by Cambridge University Press

own phonological system. Thus, both reading and repetition may obscure speakers' true production ability, which could lead to a fuzzy or inaccurate view of the perception-production link (for studies that address differences in performance after orthographic vs. auditory input, see, e.g., Davidson, 2010; de Jong et al., 2009; Kato & Baese-Berk, 2020). One controlled task that avoids both confounds is picture naming. As long as researchers take care to select high frequency, imageable items that are likely to be familiar to speakers, picture naming may provide the clearest perspective on controlled production ability.

On the other side of the controlled-spontaneous spectrum, picture description and story narration have been used to investigate spontaneous production knowledge. These tasks are advantageous for understanding what speakers actually produce under realistic speaking conditions, but they also imply some risk. For instance, participants may not produce enough tokens of the target sound for reliable analysis. Furthermore, other sources of variation in spontaneous production, such as the phonetic context in which the sounds occur and the lexical characteristics of the carrier word, must be accounted for during data analysis. These concerns can be mitigated by, for example, providing speakers with a list of target words that they should use to describe the picture (e.g., Nagle, 2021; Saito & van Poeteren, 2017), but providing written target words introduces the same potential orthographic and literacy confounds that reading tasks do.

After researchers have determined what tasks they will use to examine controlled and/or spontaneous production knowledge, they must determine how they will define accuracy. Acoustic measurements provide an objective measure, but even acoustic measurements involve an element of choice (e.g., what features to measure, how to measure them), and such measurements may not reveal much about how listeners actually process L2 speech. Moreover, L2 speakers may implement L2 contrasts in nonnativelike ways, in which case they might use phonetic cues that acoustic analyses of canonical features (i.e., the features that monolingual speakers use to perceive and produce the target sound) would not detect. For this reason, targeted acoustic analysis may not capture all the differences L2 speakers make between sounds. Furthermore, it is rarely the case that speakers use a single phonetic cue to differentiate sounds, but without complex analytical tools, it can be difficult to integrate multiple cues into a single analysis. Some recent work using linear discriminant analysis (Mairano et al., 2019) has attempted to address this problem. The challenge of an integrated analysis, however, is understanding precisely how each cue contributes to the differentiation of sounds categories in production.

Listener-based measures offer an alternative perspective on production accuracy. Whereas acoustic measurements shed light on phonetic nativelikeness, listener-based approaches are rooted in the notion of intelligibility, or the extent to which speakers can produce an intelligible L2 contrast (the word the speaker intends to produce is *beat*, and the listener perceives the word as intended, hearing *beat* instead of *bit*; see, e.g., Munro & Derwing, 2008). Intelligibility has broad ecological validity, but it is also a relatively coarse-grained measure that does not provide insight into the precise characteristics of L2 sounds, which is often one of the central goals of (perception-)production research. For instance, L2 speakers may produce a covert acoustic contrast between two L2 sounds even if listeners do not perceive a difference (Song & Eckman, 2019). Listeners can also be asked to evaluate production accuracy using a scalar rating system (see, e.g., Kissling, 2014; Lopez-Soto & Kewely -Port, 2009; Rochet, 1995; Saito & van Poeteren, 2017), or

https://doi.org/10.1017/S0272263121000371 Published online by Cambridge University Press

by comparing pre and posttest productions to one another (Bradlow et al., 1997). However, relying on listener perception to assess production accuracy can introduce perceptual biases into production measures. To that point, a wide variety of factors, including lexicality, lexical frequency, semantic plausibility, and social information, are known to influence perception, especially the perception of ambiguous sounds (e.g., Ganong, 1980).

In summary, then, researchers must take care to determine what facet of production they are interested in examining. No single production task or accuracy measure can fully capture production performance at a single time point, much less how L2 production develops over time. Thus, we advise researchers to measure production in both controlled and spontaneous speech (and examine how different views of production accuracy relate to one another), although few studies have done so to date (for notable exceptions, see Kartushina & Frauenfelder, 2014; Lambacher et al., 2005; Saito & van Poeteren, 2017; Wang et al., 2003).

EVALUATING THE PERCEPTION-PRODUCTION LINK

To evaluate the perception-production link, researchers must decide what perception and production measures should be compared and what would count as evidence of a robust link. Selecting measures for comparison is not trivial. Perception measures such as *d'* represent sensitivity to contrast, whereas production measures represent the ability to produce the target phone intelligibly or accurately. It is, therefore, unclear whether a measure that encodes information about a contrast should be related to the production of either of the phones that make up that contrast. At the most basic level, it seems important to compare like with like; measures that tap into fine-grained aspects of perception (patterns of cue use, precise boundary locations, etc.) should be paired with fine-grained production measures (e.g., Schertz et al. 2015), and categorical perception measures (discrimination and identification indices) should be paired with categorical production measures (intelligibility). It bears mentioning that intelligibility is often operationalized as the percentage of words that are correctly identified as the target word (i.e., the word the speaker intended to produce). However, intelligibility can also be quantified by means of contrast sensitivity metrics. In this case, the resulting score would represent the extent to which listeners perceive the contrast that the L2 speaker intended to produce. Such a production measure might be more closely aligned with its perceptual counterpart: *d'* for perception would represent the L2 learner's ability to discriminate contrastive L2 sounds, and *d'* for production would represent that individual's ability to produce a discriminable L2 contrast.

Another issue that deserves scrutiny is the fact that perception tests are implemented in a controlled listening context in which the target forms are often presented in isolation without noise. Yet, production can be measured in controlled or spontaneous speech. It stands to reason that perception measures would bear a stronger relationship to controlled production measures than spontaneous production measures (Nagle, 2021). One future goal for perception research should, therefore, be to test perception in more spontaneous listening contexts that require listeners to process speech for meaning (see, e.g., Kim et al., 2020).

https://doi.org/10.1017/S0272263121000371 Published online by Cambridge University Press

Evaluating perception-production links also entails making choices about statistical tests. This choice is clearly constrained by the research design and research questions, but a variety of descriptive and inferential approaches are nonetheless possible. In early research, it was common to analyze perception and production separately and subsequently compare performance in the two modalities (Bohn & Flege, 1997; Caramazza et al., 1973; Gass, 1984; Mack, 1989; Sheldon & Strange, 1982; Zampini, 1998). This analysis constitutes a type of rank ordering that reveals broad information on perception-production patterns such as accuracy in perception is better than (i.e., leads or precedes) accuracy in production, accuracy in production seems to outpace accuracy in perception, and so on. Few contemporary studies use rank ordering exclusively, but it is often presented as one aspect of a more comprehensive set of analyses. For instance, numerous studies that have rank ordered contrasts according to their difficulty in perception and production (Evans & Alshangiti, 2018; Jia et al., 2006; Kartushina & Frauenfelder, 2014; Rallo Fabra & Romero, 2012) have shown that difficulty in one modality is not always reflected in the other.

Correlation and regression are the most common inferential tests used to assess the strength of perception-production relationships. It is important to recognize that neither test suggests directionality or causality, and, when applied to cross-sectional data, these techniques run the risk of mischaracterizing the link based on the precise moment at which it was measured.

Some researchers have achieved a finer-grained perspective by using gain scores to examine whether changes in perception are associated with changes in production (Huensch & Tremblay, 2015; Kartushina & Frauenfelder, 2014). Yet, this approach has similar limitations insofar as it renders a relatively circumscribed view of what should be considered a dynamic developmental phenomenon. It comes as no surprise then, that findings range from no significant perception-production link at all in some studies (Hattori & Iverson, 2010; Kartushina & Frauenfelder, 2014; Schertz et al., 2015; Thorin et al., 2018) to medium to large correlations in others (Baker & Trofimovich, 2006; Borden et al., 1983; Lopez-Soto & Kewley-Port, 2009).

Examining the effect of (perception) training is another common means of testing the link. Rather than directly comparing performance in the two modalities (or in addition to this comparison), training studies address how experience in one modality impacts learning in the other. According to Sakai and Moorman's (2018) meta-analysis of perception training studies, perception training leads to small, yet significant, gains in posttest production accuracy. At the same time, they found that gains in perception were not significantly correlated with gains in production, and individual perception training studies have yielded a wide range of results. Researchers have also investigated the effect of production training and integrated perception-production training paradigms on perceptual learning. Here too, a range of effects have been observed: production training leading to medium gains in perception (Sakai, 2016); no influence of production training on perception (Thorin et al., 2018); and disruption of perceptual learning (Baese-Berk, 2019; Baese-Berk & Samuel, 2016). However, as in other areas of perception-production research, training studies vary widely in methodological choices, choices that have a direct impact on perception-production findings (cf. Sakai & Moorman, 2018).

Apart from these issues, current approaches suffer from two weaknesses. First, most studies have been relatively short-term, insofar as they have analyzed the link at a single

point in time or over two to three data points (e.g., pre-post-delayed training studies). Yet, a robust test of the link would require demonstrating that within-subjects changes in perception guide within-subjects changes in production consistently over longer periods. That is, perception must be conceptualized as a time-varying predictor of production (Nagle, 2021). It would also be important to test how the strength of the link changes over time, as might be accomplished through Time × Perception interaction terms. Another important topic that research has not evaluated statistically is directionality. Directionality in longitudinal studies can be evaluated through cross-lagged analyses, such as cross-lagged panel models (longitudinal structural equation models that can estimate the impact of one variable on another over time; for an introduction to longitudinal structural equation models, see Little, 2013). Such models can be used to evaluate Flege and Bohn's (2021) hypothesis that perception and production co-evolve. They can also provide insight into the weights of reciprocal relationships. One challenge will be achieving the sample size required for statistical power.

RECOMMENDATIONS FOR FUTURE WORK

Cleary, perception-production research is a complex theoretical and methodological undertaking. In this section, we synthesize general recommendations for conducting robust perception-production research and provide illustrative examples of perception-production research questions and the methods that would be needed to address them. First, as discussed elsewhere, categorical perception measures should be paired with categorical production measures, and gradient perception measures should be paired with gradient production measures. If perception is operationalized as the ability to perceive contrast, then production should be defined in similar terms as the ability to produce contrast, and if perception tasks tap into sensitivity to fine-grained phonetic distinctions, then production tasks should also tap into the ability to produce fine-grained phonetic detail. It is unclear if crossed categorical/gradient comparisons are methodologically valid (e.g., pairing a categorical perception measure with a gradient production measure), and even if they are, they may not yield meaningful findings.

To provide a concrete example, consider the case of stop consonant perception and production. In one study, we might examine whether participants' ability to identify stop-initial words produced by multiple speakers (a task that aligns well with what listeners actually do when processing speech) is related to their ability to produce an intelligible stop-consonant contrast. To quantify production, participants could be asked to name pictures, and participants' productions could then be presented to native listeners in an identification task that mirrors the perception task that participants completed. In such a study, the focus is on a categorical perception-production link given that perception and production are defined in terms of perceiving and producing contrast. In another study, we might be interested in the extent to which changes in perceptual boundaries or cue use are associated with changes in production accuracy. Here, given that the perceptual task and outcome measure are fine-grained, an equally fine-grained production outcome measure would be appropriate, such as acoustic accuracy measures or cue weights (see, e.g., Schertz et al., 2015). By taking a gradient approach to the perception-production link, this study could provide insight into whether subtle changes in the phonetic organization of perceptual categories are reflected in the accuracy with which those categories are

https://doi.org/10.1017/S0272263121000371 Published online by Cambridge University Press

phonetically realized in production. These two examples underscore the need to consider and specify in clear terms the conceptual focus of perception-production research. Whereas the first study is aligned with the notion of contrast perception and its relationship to intelligibility, the second is concerned with perceptual foreign accents and their relationship to phonetic accuracy.

We also encourage researchers to study perception-production links longitudinally. Longitudinal perception-production work brings its own set of challenges, such as selecting and sequencing tasks in a manner that will reveal how various facets of perception and production develop and interact with one another. In this area, current models provide some guidance. If, as the SLM(-r) suggests, listeners must discern differences between crosslinguistically similar sounds before they can create new phonetic categories in the L2, then it is important to examine both discrimination/categorization and identification over time using appropriate perceptual tasks. Likewise, production accuracy should be examined longitudinally because as identification accuracy improves (i.e., as new phonetic categories emerge and take shape), production should also begin to improve. Although current models make broad hypotheses on the temporal characteristics of the perception-production relationship, they do not provide information on its timing. That is, they do not specify the point at which development in one area begins to affect development in the other. Thus, it is currently unclear if discrimination accuracy must reach a certain threshold before identification accuracy will begin to improve (and if identification accuracy must reach a certain threshold before production accuracy will begin to improve), or if both skills improve simultaneously.

Longitudinal research is uniquely positioned to provide insight into these developmental questions (Nagle, 2018). At the same time, we acknowledge that longitudinal research is logistically complex and time- and cost-intensive. Given this state of affairs, it is understandable that, to the best of our knowledge, there have not been any large-scale longitudinal studies of perception-production learning. Collaboration (e.g., multi-site studies) may be the key to achieving such designs.

With respect to learning scenario, we suggest that researchers carefully consider the nature of the learning scenario, thinking about both the number of categories involved in L1 and L2 contrasts and the phonetic cues with which they are implemented. It would be advantageous to systematically compare different learning scenarios, while holding other elements of methodology constant (e.g., participant characteristics, task characteristics). Such studies have the potential to shed light on whether each learning scenario is associated with a qualitatively distinct perception-production relationship. This information can, in turn, inform training paradigms designed to maximize gains in each modality.

Finally, we would like to draw attention to a few additional conceptual considerations that are applicable to, but extend beyond the boundaries of, perception-production research. Pronunciation researchers have long acknowledged that comparing L2 speakers to monolingual native speakers is like comparing apples to oranges. Thus, it comes as no surprise that L2 pronunciation research has increasingly shifted toward a bilingual baseline (Sakai, 2018). Yet, perception-production research to date has been dominated by monolingual norms, especially with respect to acoustic analysis. On the one hand, this is sensible, because research on monolingual norms can help researchers narrow the analysis to a plausible set of phonetic cues. On the other hand, an emphasis on monolingual norms may preclude a more nuanced understanding of the complex ways in which L2

learners come to perceive and produce L2 contrasts. It is, therefore, important that pronunciation researchers, including perception-production researchers, begin to explore more robust methods of acoustic analysis that examine and/or combine multiple acoustic cues.

Thus, in perception-production research, a bilingual baseline entails examining the perception-production patterns of advanced L2 users (i.e., L2 speakers who are highly intelligible and comprehensible, although not necessarily nativelike), who can in turn serve as a meaningful point of comparison for L2 learners. Furthermore, it may be useful to reexamine whether L2 perception and production ever settle into a completely stable state. Perception and production may vary even in advanced L2 users. Dynamic approaches (see, e.g., Hiver & Al-Hoorie, 2019) may be particularly well-suited to disentangle systematic and stochastic variance in perception and production over time.

Last but not least, we encourage researchers to practice open science by making their tasks and materials publicly available whenever possible (e.g., through platforms such as OSF, IRIS). Open materials offer several advantages for perception-production research. For one, as a field, it would allow us to crowdsource the most valid and reliable tasks. Second, it would encourage replication, leading to a greater number of methodologically similar studies suitable for comparison and meta-analysis.

MODEL BUILDING

As we have outlined elsewhere, PAM (Best, 1995) and the revision focused on L2 learners (Best & Tyler, 2007), the SLM (Flege, 1995) and its revision (Flege & Bohn, 2021), and the L2LP (van Leussen & Escudero, 2015) primarily focus on questions of how the L1 and L2 interact. In this section, we briefly recap the extent to which these models address the perception-production link before turning to what a comprehensive perception-production model should address.

PAM and PAM-L2 concentrate on the perception of novel sounds in the L2. Because these models are rooted in direct realism (articulatory gestures are the basis of perception), then changes in one modality ought to mirror changes in the other. The SLM also makes predictions about the relationship between the two modalities. In the original instantiation of the model, Flege (1995) predicted that perception should precede production. However, the revised model predicts that perception and production should "co-evolve without precedence" (Flege & Bohn, 2021, p. 42). That is, the model predicts a bidirectional link between the two modalities.

The challenge for these and other models that take a similar stance is that such a wide range of experimental outcomes are consistent with all of these hypotheses. That is, correlations between perception and production could be taken as evidence for a perception-first, production-first, or coevolution hypotheses. That is, while they make testable predictions about a variety of aspects of L2 speech sound learning, the models do not make clear, testable predictions or propose falsifiable hypotheses regarding the relationship between the two modalities.

Some previous work has suggested that models from outside the perception and production domain could better explain the types of learning relationships we observe. For example, Baese-Berk (2019) proposed a shared-resources account for perception and production learning. This account builds on Ferreira & Pashler (2002), who used a central

https://doi.org/10.1017/S0272263121000371 Published online by Cambridge University Press

bottleneck theory to explain interference during word production. That is, it is possible that accounts for other aspects of speech and language production, or learning outside of the language domain, may be recruited to develop models specific to the phenomena discussed here. In fact, one important goal of any linguistic subdomain is to contribute findings that can inform theories of language and cognition.

We suggest here that the field is ready for new models (or modules within existing models) that explicitly account for and make testable predictions about the relationship between the two modalities. We propose that such models should consider learning scenarios, drawing from previous models. We also propose that models should consider development, predicting how perception-production relationships may shift over time, and they should make explicit predictions about what types of behaviors we would expect on different tasks at different points in development. That is, given that perception and production can be assessed and compared in a variety of ways, we believe a comprehensive model accounting for the relationship between perception and production must make clear, testable predictions about behavior across a range of tasks, learning scenarios, and developmental stages. This is not a trivial undertaking, but using existing models that were not designed to account for this relationship is no longer the best way of advancing research in this area.

CONCLUSION

Research on the relationship between L2 speech perception and L2 speech production has yielded diverse, and at times seemingly irreconcilable, findings ranging from dissociations to large correlations. We believe that this diversity reflects the diverse conceptual and methodological choices that researchers have made. Furthermore, current models of L2 sound learning provide some insight into perception-production relationships, but it is important to bear in mind that they were not designed to explain how perception and production interact. Rather, they were developed to account for L1-L2 interactions in specific learner populations. As a result, the broad claims that they make are often open to a range of interpretations and research methodologies. In this article, we have surveyed theoretical issues that perception-production research must address to derive specific, testable perception-production hypotheses. We have also surveyed and made recommendations for conducting methodologically robust perception-production research. Ultimately, the time is right to take a step back from current models, revisiting the assumptions we make about perception-production links in L2 sound learning and the methods we use to test them.

COMPETING INTERESTS

The authors declare none.

NOTES

[1] We acknowledge that it would also be important to examine the perception and production of suprasegmental features. However, it is unclear if suprasegmentals involve the same type of category learning as segmentals. Many suprasegmental features are far more gradient in nature and serve a paralinguistic function.

https://doi.org/10.1017/S0272263121000371 Published online by Cambridge University Press

There is also a lack of research on perception-production relationships for suprasegmental features, which means that it would not be possible at this time to examine the state of the scholarship in that area. This is a topic that future research should address, especially because coordinating segmental and suprasegmental perception-production research can lead to more comprehensive models of speech learning.

[2]Because naturally produced tokens vary along a variety of phonetic dimensions, it is also unclear precisely what information listeners are using to make their perceptual judgments. Thus, such stimuli lend themselves well to outcome measures that reflect sensitivity to contrast, rather than outcome measures designed to provide insight into the precise characteristics of phonetic categories.

[3]The inclusion of orthography can introduce an additional confound related to literacy. Poor performance on perceptual tasks involving orthographic matches could be due to an incomplete understanding of novel phoneme-to-grapheme representations, not an inability to perceptually identify a target sound.

[4]ABX tasks are also sometimes conducted as AXB tasks where the token inducing the response is presented between the two anchor stimuli instead of after them.

[5]Accounting for response bias is important because individuals may show a preferred response pattern that can muddy findings. Moreover, experimental design decisions (stimuli, instructions, and so on) can also influence response bias.

REFERENCES

Baese-Berk, M. M. (2019). Interactions between speech perception and production during learning of novel phonemic categories. *Attention, Perception, & Psychophysics*, *81*, 981–1005. https://doi.org/10.3758/s13414-019-01725-4

Baese-Berk, M. M. (2010). *An examination of the relationship between speech perception and production* (Publication No. 3433556) [Doctoral dissertation, Northwestern University]. ProQuest Dissertations & Theses Global.

Baese-Berk, M. M., & Samuel, A. G. (2016). Listeners beware: Speech production may be bad for learning speech sounds. *Journal of Memory and Language*, *89*, 23–36. https://doi.org/10.1016/j.jml.2015.10.008

Baker, W., & Trofimovich, P. (2006). Perceptual paths to accurate production of L2 vowels: The role of individual differences. *IRAL – International Review of Applied Linguistics in Language Teaching*, *44*, 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: Issues in cross-language research* (pp. 171–204). York Press.

Best, C. T., & Tyler, M. D. (2007). Nonnative 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. 13–24). John Benjamins.

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

Borden, G., Gerber, A., & Milsark, G. (1983). Production and perception of the /r/-/l/ contrast in Korean adults learning English. *Language Learning*, *33*, 499–526. https://doi.org/10.1111/j.1467-1770.1983.tb00946.x

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, *101*, 2299–2310. https://doi.org/10.1121/1.418276

Caramazza, A., Yeni-Komshian, G. H., Zurif, E. B., & Carbone, E. (1973). The acquisition of a new phonological contrast: The case of stop consonants in French-English bilinguals. *Journal of the Acoustical Society of America*, *54*, 421–428. https://doi.org/10.1121/1.1913594

Casillas, J. V. (2020a). The longitudinal development of fine-phonetic detail: Stop production in a domestic immersion program. *Language Learning*, *70*, 768–806. https://doi.org/10.1111/lang.12392

Casillas, J. V. (2020b). Phonetic category formation is perceptually driven during the early stages of adult L2 development. *Language and Speech*, *63*, 550–581. https://doi.org/10.1177/0023830919866225

Cerviño-Povedano, E., & Mora, J. C. (2010). *Investigating Catalan-Spanish bilingual EFL learners' over-reliance on duration: Vowel cue weighting and phonological short-term memory*. In K. Dziubalska–Kołaczyk, M. Wrembel, & M. Kul (Eds.), Achievements and perspectives in the acquisition of second language speech: New sounds *2010* (pp. 53–64). Peter Lang.

Chandrasekaran, B., Koslov, S. R., & Maddox, W. T. (2014). Toward a dual-learning systems model of speech category learning. *Frontiers in Psychology*, *5*, 1–17. https://doi.org/10.3389/fpsyg.2014.00825

Darcy, I., Mora, J. C., & Daidone, D. (2016). The role of inhibitory control in second language phonological processing. *Language Learning*, *66*, 741–773. https://doi.org/10.1111/lang.12161

Davidson, L. (2010). Phonetic bases of similarities in cross-language production: Evidence from English and Catalan. *Journal of Phonetics*, *38*, 272–288. https://doi.org/10.1016/j.wocn.2010.01.001

de Jong, K., Hao, Y. C., & Park, H. (2009). Evidence for featural units in the acquisition of speech production skills: Linguistic structure in foreign accent. *Journal of Phonetics*, *37*, 357–373. https://doi.org/10.1016/j.wocn.2009.06.001

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

Escudero, P. (2007). Second-language phonology: The role of perception. In M. C. Pennington (Ed.), *Phonology in context* (pp. 109–134). Palgrave Macmillan.

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

Ferreira, V. S., & Pashler, H. (2002). Central bottleneck influences on the processing stages of word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 1187–1199. https://doi.org/10.1037/0278-7393.28.6.1187

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

Flege, J. E., & Bohn, O.-S. (2021). The revised Speech Learning Model (SLM-r). In R. Wayland (Ed.), *Second language speech learning: Theoretical and empirical progress* (pp. 3–83). Cambridge University Press.

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

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, *14*, 3–28. https://doi.org/10.1016/S0095-4470(19)30607-2

Gambi, C., & Pickering, M. J. (2013). Prediction and imitation in speech. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00340

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*, 110–125. https://doi.org/10.1037/0096-1523.6.1.110

Gass, S. (1984). Development of speech perception and speech production abilities in adult second language learners. *Applied Psycholinguistics*, *5*, 51–74. https://doi.org/10.1017/S0142716400004835

Hattori, K., & Iverson, P. (2010). *Examination of the relationship between L2 perception and production: An investigation of English /r/-/l/ perception and production by adult Japanese speakers*. *Paper presented at the Second Language Studies: Acquisition*, Learning, Education and Technology, Tokyo.

Hiver, P., & Al-Hoorie, A. (2019). *Research methods for complexity theory in applied linguistics*. Multilingual Matters.

Huensch, A., & Tremblay, A. (2015). Effects of perceptual phonetic training on the perception and production of second language syllable structure. *Journal of Phonetics*, *52*, 105–120. https://doi.org/10.1016/j.wocn.2015.06.007

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

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*, 1–17. 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. *Journal of the Acoustical Society of America*, *138*, 817–832. https://doi.org/10.1121/1.4926561

Kato, M., & Baese-Berk, M. M. (2020). The effect of input prompts on the relationship between perception and production of non-native sounds. *Journal of Phonetics*, *79*, 100964. https://doi.org/10.1016/j.wocn.2020.100964

Kim, J.-e., Cho, Y., Cho, Y., Hong, Y., Kim, S., & Nam, H. (2020). The effects of L1–L2 phonological mappings on L2 phonological sensitivity. *Studies in Second Language Acquisition*, *42*, 1041–1076. https://doi.org/10.1017/s0272263120000133

Kissling, E. M. (2014). What predicts the effectiveness of foreign-language pronunciation instruction? Investigating the role of perception and other individual differences. *Canadian Modern Language Review*, *70*, 532–558. https://doi.org/10.3138/cmlr.2161

Lambacher, S. G., Martens, W. L., Kakehi, K., Marasinghe, C. A., & Molholt, G. (2005). The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics*, *26*, 227–247. https://doi.org/10.1017/S0142716405050150

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*, 384–422. https://doi.org/10.1080/00437956.1964.11659830

Little, T. (2013). *Longitudinal structural equation modeling*. The Guilford Press.

Lopez-Soto, T., & Kewley-Port, D. (2009). Relation of perception training to production of codas in English as a second language. *Proceedings of Meetings on Acoustics* (Vol. *6*). Acoustical Society of America.

Mack, M. (1989). Consonant and vowel perception and production: Early English-French bilinguals and English monolinguals. *Perception & Psychophysics*, *46*, 187–200.

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 *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 1104–1108).

Mayr, R., & Escudero, P. (2010). Explaining individual variation in L2 perception: Rounded vowels in English learners of German. *Bilingualism: Language and Cognition*, *13*, 279–297. https://doi.org/10.1017/S1366728909990022

Munro, M. J., & Derwing, T. M. (2008). Segmental acquisition in adult ESL learners: A longitudinal study of vowel production. *Language Learning*, *58*, 479–502. https://doi.org/10.1111/j.1467-9922.2008.00448.x

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

Nagle, C. (2021). Revisiting perception-production relationships: Exploring a new approach to investigate perception as a time-varying predictor. *Language Learning*, *71*, 243–279. https://doi.org/10.1111/lang.12431

Perrachione, T. K., Lee, J., Ha, L. Y., & Wong, P. C. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *Journal of the Acoustical Society of America*, *130*, 461–472. https://doi.org/10.1121/1.3593366

Plonsky, L. (Ed.) (2015). *Advancing quantitative methods in second language research*. Routledge.

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

Rochet, B. L. (1995). Perception and production of second-language speech sounds by adults. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 379–410). York Press.

Sakai, M. (2016). *(Dis)connecting perception and production: Training adult native speakers of Spanish on the English /i/-/ɪ/ distinction* (Publication No. 10250896) [Doctoral dissertation, Georgetown University]. ProQuest Dissertations & Theses Global.

Sakai, M. (2018). Moving towards a bilingual baseline in second language phonetic research. *Journal of Second Language Pronunciation*, *4*, 11–45. https://doi.org/10.1075/jslp.00002.sak

Sakai, M., & Moorman, C. (2018). Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research. *Applied Psycholinguistics*, *39*, 187–224. https://doi.org/10.1017/S0142716417000418

Saito, K. (2013). Age effects on late bilingualism: The production development of /ɹ/ by high-proficiency Japanese learners of English. *Journal of Memory and Language*, *69*, 546–562. https://doi.org/10.1016/j.jml.2013.07.003

Saito, K. (2019). Individual differences in second language speech learning in classroom settings: Roles of awareness in the longitudinal development of Japanese learners' English /ɹ/ pronunciation. *Second Language Research*, *35*, 149–172. https://doi.org/10.1177%2F0267658318768342

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

Saito, K., & van Poeteren, K. (2017). The perception-production link revisited: The case of Japanese learners' English /ɹ/ performance. *International Journal of Applied Linguistics*, *28*, 3–17. https://doi.org/10.1111/ijal.12175

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

Schouten, B., Gerrits, E., & Van Hessen, A. (2003). The end of categorical perception as we know it. *Speech Communication*, *41*, 71–80. https://doi.org/10.1016/S0167-6393(02)00094-8

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*, 243–261. https://doi.org/10.1017/S0142716400001417

Song, J. Y., & Eckman, F. (2019). Covert contrasts in the acquisition of English high front vowels by native speakers of Korean, Portuguese, and Spanish. *Language Acquisition*, *26*, 436–456. https://doi.org/10.1080/10489223.2019.1593415

Stölten, K., Abrahamsson, N., & Hyltenstam, K. (2014). Effects of age and speaking rate on voice onset time. *Studies in Second Language Acquisition*, *37*, 71–100. https://doi.org/10.1017/s0272263114000151

Thorin, J., Sadakata, M., Desain, P., & McQueen, J. M. (2018). Perception and production in interaction during non-native speech category learning. *Journal of the Acoustical Society of America*, *144*, 92. https://doi.org/10.1121/1.5044415

van Leussen, J.-W., & Escudero, P. (2015). Learning to perceive and recognize a second language: the L2LP model revised. *Frontiers in Psychology*, *6*, 1–12. https://doi.org/10.3389/fpsyg.2015.01000

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

Werker, J. F., & Logan, J. S. (1985). Cross-language evidence for three factors in speech perception. *Perception & Psychophysics*, *37*, 278–286. https://doi.org/10.3758/BF03207136

Zampini, M. (1998). The relationship between the production and perception of L2 Spanish stops. *Texas Papers in Foreign Language Education*, *3*, 85–100.