MDPI

*Article*

# Perceived Phonological Overlap in Second-Language Categories: The Acquisition of English /r/ and /l/ by Japanese Native Listeners

**Michael D. Tyler**

School of Psychology and the MARCS Institute for Brain, Behavior and Development, Western Sydney University, Penrith, NSW 2751, Australia; m.tyler@westernsydney.edu.au

**Abstract:** Japanese learners of English can acquire /r/ and /l/, but discrimination accuracy rarely reaches native speaker levels. How do L2 learners develop phonological categories to acquire a vocabulary when they cannot reliably tell them apart? This study aimed to test the possibility that learners establish new L2 categories but perceive phonological overlap between them when they perceive an L2 phone. That is, they perceive it to be an instance of more than one of their L2 phonological categories. If so, improvements in discrimination accuracy with L2 experience should correspond to a reduction in overlap. Japanese native speakers differing in English L2 immersion, and native English speakers, completed a *forced category goodness rating task*, where they rated the goodness of fit of an auditory stimulus to an English phonological category label. The auditory stimuli were 10 steps of a synthetic /r/–/l/ continuum, plus /w/ and /j/, and the category labels were L, R, W, and Y. Less experienced Japanese participants rated steps at the /l/-end of the continuum as equally good versions of /l/ and /r/, but steps at the /r/-end were rated as better versions of /r/ than /l/. For those with more than 2 years of immersion, there was a separation of goodness ratings at both ends of the continuum, but the separation was smaller than it was for the native English speakers. Thus, L2 listeners appear to perceive a phonological overlap between /r/ and /l/. Their performance on the task also accounted for their responses on /r/–/l/ identification and AXB discrimination tasks. As perceived phonological overlap appears to improve with immersion experience, assessing category overlap may be useful for tracking L2 phonological development.

**Keywords:** Perceptual Assimilation Model; second language speech learning; English /r/ and /l/; Japanese; English as a second language; categorical perception; speech perception

## 1. Introduction

Adult second language (L2) learners almost invariably speak with a recognizable foreign accent (Flege et al. 1995; Flege et al. 1999). Less obvious for the casual observer is that they are also likely to have difficulty discriminating certain pairs of phonologically contrasting phones in the target language—that is, they also *hear* with an accent (Jenkins et al. 1995). Research into cross-language speech perception by naïve listeners has shown that attunement to the native language affects the discrimination of pairs of phonologically contrasting non-native phones (e.g., Best et al. 2001; Polka 1995; Tyler et al. 2014b; Werker and Logan 1985), often resulting in poor discrimination when both non-native phones are perceived as the same native phonological category.

For learners of an L2, the question is whether and to what extent they are able to overcome their perceptual accent and acquire new phonological categories. Discrimination of initially difficult contrasts, such as English /r/–/l/ for Japanese native listeners, can improve with naturalistic exposure (MacKain et al. 1981) and laboratory training (Bradlow et al. 1999; Bradlow et al. 1997; Lively et al. 1993; Lively et al. 1992; Lively et al. 1994; Logan et al. 1991; Shinohara and Iverson 2018). For example, in Logan et al. (1991), learners identified minimal-pair words containing /r/ or /l/ (e.g., *rake* or *lake*) and were

given corrective feedback for incorrect responses. After 15 training sessions of 40 min each, learners' identification showed a small but significant improvement. The results of the series of experiments showed that: (a) the learning generalized to novel talkers and tokens; (b) the perceptual training resulted in improved production, and; (c) improvements in both perception and production were still evident when learners were tested months later. However, in spite of the improvements from high variability training, the learners' performance did not reach the same level of accuracy as that of native speakers of English.

The fact that learners were able to identify words according to whether they contained /r/ or /l/ suggests that they had some sensitivity to the phonetic characteristics that define English /r/ versus /l/. To establish separate lexical entries for /r/–/l/ minimal pairs (e.g., *rock* and *lock*), it could be argued that they must have developed separate phonological categories for /r/ and /l/. If this is the case, it remains to be explained how they might have learned a phonological distinction without being able to discriminate it to the same level as a native speaker. The purpose of this paper is to propose a learning scenario that may account for these observations. L2 learners may, under certain learning circumstances, develop L2 phonological categories that correspond to similar sets of phonetic properties. That is, when they encounter an L2 phone, they may perceive a phonological overlap, such that the L2 phone is consistent with more than one L2 phonological category. This proposal will be tested using perception of English /r/ and /l/ by native speakers of Japanese, which has a long history of investigation. As this idea emerges from recent developments in the Perceptual Assimilation Model (PAM, Best 1994a, 1995), it complements the theoretical framework of its extension to L2 learning, the Perceptual Assimilation Model of Second Language Speech Learning (PAM-L2, Best and Tyler 2007). Thus, this chapter will present up-to-date reviews of both PAM and PAM-L2, and these will be followed by a review of Japanese listeners' perception of /r/ and /l/, and an outline of the present study.

### 1.1. The Perceptual Assimilation Model

The Perceptual Assimilation Model (PAM) was devised to account for the influence of first language (L1) attunement, both on infants' discrimination of native and non-native contrasts (Best 1994a; Best and McRoberts 2003; Tyler et al. 2014b) and on adults' discrimination of previously unheard non-native contrasts (PAM, Best 1994b; Best 1995; Best et al. 2001; Best et al. 1988; Faris et al. 2018; Tyler et al. 2014a). Following the ecological theory of perception (e.g., Gibson 1966, 1979), PAM proposes that articulatory gestures are perceived directly. Phonological categories are the result of perceptual learning, where the perceiver tunes into the higher-order invariant properties that define the category across a range phonetic variability. While they are abstract, in the sense of being coarser grained than specific articulatory movements, or the acoustic structure that corresponds to those movements, phonological categories are perceptual units rather than mental constructs (Goldstein and Fowler 2003). According to PAM, discrimination accuracy for non-native phones depends on whether and how they are *assimilated* to the native phonological system. When adults encounter non-native contrasts, their natively tuned perception may sometimes help perception (e.g., when the non-native phones are assimilated to two different native categories) and it may sometimes hinder perception (e.g., when both non-native phones are assimilated to the same native category).

An individual non-native phone can be perceptually assimilated as either *categorized* or *uncategorized*, or not assimilated to speech (Best 1995). Categorized phones are perceived as good, acceptable, or deviant versions of a native phonological category. Uncategorized phones are those that are perceived as speech, but not as any one particular native category, and non-assimilable phones are perceived as non-speech. For example, Zulu clicks were often perceived by native English speakers as finger snaps or claps (Best et al. 1988).

The perceptual learning that shapes phonological categories is driven not only by detecting the higher-order phonetic properties that define category membership, but also those set a category apart from other phonological categories in the system (see Best 1994b, pp. 261–62). PAM predicts relative discrimination accuracy for pairs of phonologically

contrasting non-native phones according to how each one is assimilated to the native phonological space. That is, discrimination accuracy for never-before-heard non-native phones depends, to a large extent, on whether the listener detects a *native* phonological distinction between the non-native phones. Consider, first, those contrasts where both non-native phones are assimilated to a native category. If each phone is assimilated to a different category (a *two-category* assimilation), then a native phonological distinction will be detected, and discrimination will be excellent. When both contrasting non-native phones are assimilated to the same native category, there is no native phonological distinction to detect and discrimination depends on whether the listener detects a difference in phonetic goodness of fit to the same native category.[1] If there is a goodness difference (a *category-goodness* assimilation), then discrimination will be moderate to very good, depending on the magnitude of the perceived difference, and if there is no goodness difference (a *single-category* assimilation), then discrimination will be poor. Thus, for discrimination accuracy, PAM predicts: two category > category goodness > single category. This main PAM prediction has been confirmed in studies on non-native consonant (Antoniou et al. 2012; Best et al. 2001) and vowel perception (Tyler et al. 2014a).

Turning to contrasts involving uncategorized phones, Best (1995) suggested that when one phone is uncategorized and the other is categorized (an *uncategorized-categorized* assimilation), discrimination should be very good. When both are uncategorized (an *uncategorized-uncategorized* assimilation), discrimination should vary depending on their phonetic proximity to each other. Faris et al. (2016) expanded on this description by describing the different ways that uncategorized non-native phones might be assimilated to the native phonological space. On each trial of a categorization task, Egyptian-Arabic listeners assigned a native orthographic label to an Australian-English vowel (in a /hVbə/ context, where V denotes the target vowel) and then rated its goodness-of-fit. To rule out random responding, Faris et al. tested whether each label was selected above chance for each vowel. An uncategorized native phone was deemed to be *focalized* if only one label was selected above chance (but below the categorization threshold, usually 50% or 70%, Antoniou et al. 2013; Bundgaard-Nielsen et al. 2011b), *clustered* if more than one label was selected above chance, or *dispersed* if no label was selected above chance. It is important to note that, by definition, listeners recognize weak similarity to native phonological categories when non-native phones are perceived as focalized or clustered, but no native phonological similarity when they are perceived as dispersed.

These expanded assimilation types lead to new predictions for uncategorized-categorized and uncategorized-uncategorized assimilations. Faris et al. (2018) suggested that the discrimination of contrasts involving focalized and clustered assimilations should vary according to the overlap in the set of categories that are consistent with one non-native phone versus the set that is consistent with the other phone. This is known as *perceived phonological overlap* (see also, Bohn et al. 2011). For example, if both non-native phones were clustered, and they were weakly consistent with the same set of native phonological categories, then they would be completely overlapping. Discrimination for completely overlapping contrasts should be less accurate than for contrasts that are only partially overlapping, and most accurate for non-overlapping contrasts (i.e., those where unique sets of labels are chosen for each non-native phone). To test this prediction, Australian-English listeners completed categorization with goodness rating and AXB discrimination tasks with Danish vowels. In AXB discrimination, participants were presented with three different vowel (V) tokens (in a /hVbə/ context) and asked to indicate whether the middle element (X) was the same as the first (A) or third (B) element. While Faris et al. did not observe any completely overlapping contrasts, they demonstrated that non-overlapping contrasts were discriminated more accurately than partially overlapping contrasts. This shows that naïve listeners are influenced by perceptual assimilation to the native phonological space even when a non-native phone is perceived as weakly consistent with one or more native

---

[1]  See Tyler (2021) for a discussion of the different sources of information that might be available for the discrimination of non-native contrasts.

phonological categories. The purpose of this study is to demonstrate how perceived phonological overlap might account for speech perception in learners who have acquired new L2 phonological categories, which is the focus of PAM's extension to L2 speech learning, PAM-L2.

*1.2. PAM-L2*

The Perceptual Assimilation Model of Second Language Speech Learning (PAM-L2, Best and Tyler 2007) was devised for predicting the likelihood of new category formation when a learner acquires an L2. Taking the naïve perceiver described by PAM as its starting point, learners are assumed to accommodate phonological categories from all of their languages in a common interlanguage phonological system. Phonological categories may be shared between the L1 and L2 (and subsequently learned languages), or the learner may establish new L2 categories. Importantly, and in contrast to the Speech Learning Model (Flege 1995, 2003), PAM-L2 proposed that learners could maintain common L1-L2 phonological categories with language-specific L1 and L2 phonetic categories, in a similar way that allophones of a phoneme might be thought to correspond to a single phonological category. For example, an early-sequential Greek-English bilingual's common L1-L2 phonological /p/ category could incorporate language-specific phonetic variants for long-lag aspirated English [pʰ] and short-lag unaspirated Greek [p] (Antoniou et al. 2012).

PAM-L2's predictions for new L2 phonological category formation are made on the basis of PAM contrast assimilation types. To illustrate the PAM-L2 principles, Best and Tyler (2007) outlined a number of hypothetical scenarios involving a previously naïve listener acquiring an L2 in an immersion context. In the case of a two-category assimilation, the learner comes already equipped with the ability to detect the phonological difference between the non-native phonemes, through attunement to the L1. Discrimination of two-category contrasts is predicted to be excellent at the beginning of L2 acquisition and the learner would develop a common L1/L2 phonological category for each of the non-native phonemes. No further learning would be required for that contrast, but L2 perception would be more efficient if they developed new phonetic categories for the L2 pronunciations of their common L1/L2 phonological categories. For single-category assimilations, the learner is unlikely to establish a new phonological category for either phone and discrimination will remain poor. In fact, both L2 phonemes are likely to be incorporated into the same L1-L2 phonological category and contrasting words in the L2 that employ those phonemes should remain homophonous for the L2 learner. For uncategorized-categorized assimilations, new L2 phonetic and phonological categories are likely to be established for the uncategorized phone, with the likelihood being higher for non-overlapping and partially overlapping assimilations than for completely overlapping assimilations (Tyler 2019).

Perhaps the most interesting case is category-goodness assimilation. According to PAM-L2, the learner is likely to develop L2 phonetic and phonological categories for the more deviant phone of the contrast. Best and Tyler (2007) speculated that the learner would first establish a new phonetic category for the more deviant phone. Initially, the deviant phone would simply be a phonetic variant of the common L1-L2 phonological category, but as the learner came to recognize that the phonetic difference between the phones signaled an L2 phonological contrast, a new L2 phonological category would emerge for the newly developed phonetic category. This new L2 phonological category would support the development of an L2 vocabulary that maintains a phonological distinction between minimally contrasting words.

Best and Tyler (2007) suggested that this perceptual learning might occur fairly early in the learning process for adult L2 acquisition. Learners with 6–12 months of immersion experience were considered to be "experienced". An increasing vocabulary was seen as a limiting influence on perceptual learning, but more recent work has shown that vocabulary size may assist learners in establishing L2 categories (Bundgaard-Nielsen et al. 2012; Bundgaard-Nielsen et al. 2011a, 2011b). It is possible that vocabulary expansion

constrains perceptual learning in the case of a single category assimilation, but facilitates perceptual learning when there is a newly established L2 phonological category (i.e., for category goodness assimilations and those involving an uncategorized phone). For example, once new L2 phonological categories are established, learners could use lexically guided perceptual retuning (McQueen et al. 2012; Norris et al. 2003; Reinisch et al. 2013) to accommodate to the phonetic properties of the newly acquired category.

The scenarios outlined in Best and Tyler (2007) assumed an idealized situation where an adult learner with no previous L2 experience is immersed in the L2 environment and where L2 input is entirely through the spoken medium. However, the majority of learners do not acquire an L2 solely from spoken language—the L2 is often acquired first in a formal learning situation using both written and spoken language, and this may occur in the learner's country of origin from a teacher who may speak the L2 with a foreign accent. While category formation could occur from the bottom up in an immersion scenario, the classroom learning environment may expose learners to information about an L2 phonological contrast before they have had the opportunity to attune to the phonetic properties necessary to discriminate it. When the L2 has a phonographic writing system, the most likely source of information would be from orthographic representations of words (see Tyler 2019, for a discussion of how orthography might influence category acquisition in the classroom).

The classroom learning environment would not change the PAM-L2 predictions for contrasts that were initially two-category assimilations, and the uncategorized phone of an uncategorized-categorized assimilation may even benefit from more rapid acquisition under those circumstances (Tyler 2019). However, for those contrasts where both L2 phones are assimilated to the same L1 category, single-category and category-goodness assimilations, the language learning environment could have profound effects on attunement. In fact, this situation may result in a new type of scenario that was not considered in Best and Tyler (2007). Let us reconsider the case of a category-goodness assimilation under those circumstances. The L1 category to which both L2 phonemes are assimilated would form a common L1-L2 phonological category with the more acceptable L2 phoneme, as in the immersion case. For the more deviant phone, rather than discovering a new L2 phonological category on the basis of attunement to articulatory-phonetic information, as was proposed for the immersion context, the learners could possibly discover the L2 phonological contrast via other sources (e.g., via orthography when it has unambiguous grapheme-phoneme correspondences). Since phonological categories are perceptual units, according to PAM-L2, the learners would need to establish a new L2 phonological category for the more deviant phone to acquire L2 words that preserved the phonological contrast. If they had not yet tuned into the phonetic differences that signal the phonological contrast in the L2, then the new L2 phonological category would correspond to a similar set of phonetic properties as the common L1-L2 category. In that situation learners may continue to have difficulty discriminating the L2 contrast, not because the two phones are assimilated to the same phonological category, but because both phones are perceived as being instances of the same two phonological categories. Just as Faris et al. (2016, 2018) have shown that a pair of non-native phones may fall in a region of phonetic space that corresponds to the same set of L1 phonological categories (i.e., a completely overlapping uncategorized-uncategorized assimilation), it is proposed here that a pair of L2 phones might be perceived as consistent with the same two L2 phonological categories.

The aim of this study is to test whether L2 learners who have acquired English in the classroom prior to immersion perceive phonological overlap between L2 phonological categories for contrasting L2 phones that were likely to have been initially perceived as category-goodness assimilations. Furthermore, with immersion experience, the categories should start to separate, such that learners who have recently arrived in an English-speaking country should exhibit greater category overlap than L2 users who have been living in the L2 environment for a long period (MacKain et al. 1981).

### 1.3. Perception of English /r/ and /l/ by Japanese Native Listeners

The learner group for the present study will be Japanese native speakers and the contrast to be tested will be the English /r/–/l/ contrast. Although the /r/–/l/ contrast was originally thought to be a single-category assimilation (Best and Strange 1992), there is now widespread agreement that it is a category-goodness assimilation (Aoyama et al. 2004; Guion et al. 2000; Hattori and Iverson 2009), with /l/ being rated as a more acceptable version of the Japanese /ɾ/ category than is English /r/. Using this contrast allows the results to be interpreted in light of the rich history of investigations into that listener group/contrast combination.

One of the earliest investigations into Japanese identification and discrimination of /r/ and /l/ was conducted by Goto (1971). He established that Japanese native speakers learning English had difficulty identifying and discriminating minimal-pair words, such as *play* and *pray* and concluded that /r/–/l/ pronunciation difficulties were likely to be perceptual in origin. Miyawaki et al. (1975) had participants discriminate steps on a synthetic /r/–/l/ continuum, in which F1 and F2 were held constant and F3 varied. Unlike native English speakers, Japanese native speakers living in Japan with at least 10 years of formal English language training did not show a categorical peak in discrimination, suggesting that they did not perceive a categorical boundary between /r/ and /l/. Interestingly, the Japanese listeners performed similarly to native English listeners when they were presented with a non-speech continuum that contained only the F3 transition (F1 and F2 values were set to zero). Thus, the Japanese listeners were able to detect the differences in frequencies along the F3 continuum, but when the same acoustic patterns were presented in a speech context, they failed to discriminate them.

In contrast to the findings of Miyawaki et al. (1975), which tested discrimination only, Mochizuki (1981) found that a group of Japanese listeners residing in the USA split the continuum into two separate categories in an /r/–/l/ identification task. This naturally led to the hypothesis that Japanese native speakers can acquire the /r/–/l/ distinction given sufficient naturalistic exposure. To test this, MacKain et al. (1981) directly compared the identification and discrimination of an /r/–/l/ continuum by Japanese native speakers who differed in their exposure to English conversation. Both groups were living in the USA at the time of testing; one group had received extensive English conversation training from a native speaker of US English whereas the other group had received little or no native English conversation training. To optimize the possibility that the less experienced group might discriminate the stimuli, MacKain et al. enriched the /r/–/l/ continuum by providing multiple redundant cues. Whereas Miyawaki et al. held F1 and F2 constant while varying F3, the stimuli of MacKain et al. varied along all three dimensions. In spite of the redundant cues, the less experienced group showed no evidence of categorical perception and their discrimination was close to chance. The more experienced group, on the other hand, perceived the stimuli categorically, and in a similar way to native US-English speakers, and although their discrimination was less accurate than the native speakers, the shape of the response function was similar. They concluded that it was indeed possible for Japanese native speakers to acquire categorical perception of English /r/ and /l/ that approximates that of native speakers.

An individual-differences approach to identification of /r/ and /l/ by Japanese native speakers was undertaken by Hattori and Iverson (2009). Their 36 participants varied in age (19 to 48 years), amount of formal English instruction (7 to 25 years), and the amount of time spent living in an English-speaking country (1 month to 13 years, with a median of 3 months). Participants completed two types of identification task. One was an /r/–/l/ identification task, in which participants heard minimal-pair English words, such as *rock* and *lock*, and indicated whether they began with /r/ or /l/. There was a wide range of accuracy, from close to chance to 100% correct, and a mean of 67% correct. The other identification task was a "bilingual" task, with consonant-vowel syllables formed from the combination of five vowels and three consonants (English /r/, English /l/, and Japanese /ɾ/). Participants were asked to indicate whether the first sound was R, L, or Japanese

R. The identification of /r/ was quite accurate, at 82% correct, and it was almost never confused with /ɾ/ (2%). Identification was less accurate for /l/ (58% correct), with errors split evenly between /r/ and /ɾ/. Japanese /ɾ/ was identified reasonably accurately (77% correct), with most confusions occurring with /l/ (16%). Clearly, the fewest confusions occurred between /r/ and /ɾ/, and the most confusion occurred for /l/ with both /r/ and /ɾ/. While the authors suggested that /l/ appears to be assimilated to Japanese /ɾ/, they did not observe a correlation between /r/–/l/ identification accuracy and the degree of confusion between /ɾ/ and /r/. That is, those who performed poorly on /r/–/l/ identification appear not to have done so because they assimilated both /r/ and /l/ to Japanese /ɾ/. Instead, they suggested that that the learners had an /r/ category that was not optimally tuned to English. It is possible that those results could be explained by perceived phonological overlap. The results as a whole are consistent with the idea being proposed here that the participants had developed a new L2 phonological category for /r/, with a corresponding English [ɹ] phonetic category, and that /l/ had assimilated to /ɾ/ to form a common L1-L2 phonological category, with language-specific [l] and [ɾ] phonetic variants.[2] Variability in identification for /r/, /l/, and /ɾ/ could be explained by the different patterns of phonetic and phonological overlap between the categories.

### 1.4. The Present Study

The aim of this study is to test the idea that learners who had acquired English in a formal learning setting may perceive phonological overlap when they encounter L2 phones, and that the overlap decreases with immersion experience in an L2 environment. Faris et al. (2018) determined overlap in L1 cross-language speech perception using a categorization task with goodness rating. Across the sample, if the same set of categories was selected above chance for two non-native vowels then the contrast was deemed to be overlapping. While this approach has been shown to provide a reliable indication of overlap in cross-language speech perception, particularly for vowels where categorization is more variable than it is for consonants, there is an assumption built into the categorization task that an individual only perceives one phonological category in the stimulus. If the participant perceives the stimulus as consistent with more than one category, then the categorization task may only provide an imperfect approximation of the amount of overlap. For example, if the stimulus is clearly more acceptable as one category than another, then the participant may only ever select the best-fitting category, and the task would fail to reveal perceived phonological overlap.

For the purposes of this study, a task is required that can identify perceived phonological overlap, for a given L2 phone, and that is capable of detecting differences in the amount of overlap between groups. This can be achieved by eliminating the categorization stage and simply rating the goodness of fit of each stimulus to a phonological category that is provided on each trial—a *forced category goodness rating task*. For example, participants would be presented with an English category label, e.g., "R as in ROCK", and an auditory stimulus, and they would be asked to rate the goodness of fit of the auditory stimulus to the given category. That task will be used in the present study to assess category overlap in Japanese learners of English with more versus less experience in an immersion environment, and in a native English control group.

The stimuli will be the 10-step rock-lock continuum, developed at Haskins Laboratories (Best and Strange 1992; Hallé et al. 1999; MacKain et al. 1981), that uses multiple redundant cues for F1, F2, and F3. The /w/ and /j/ end points of the wock-yock continuum will also be included as control stimuli, as participants are expected to indicate that those stimuli are not similar to either /r/ or /l/. Participants will rate the goodness of fit of each stimulus to four English phonological categories: /l/, /r/, /w/, and /j/. Results from across the continuum will give insight into the internal structure of the phonological category, at least along one axis of acoustic-phonetic variability, and they will also pro-

---

[2] Note that square brackets are used here to denote a phonetic category rather than a specific phone.

vide a link to previous studies that have used those stimuli. To further assist with such comparisons, participants will complete discrimination and /r/–/l/ identification tasks in addition to the forced category goodness rating task. In line with those previous studies, native English listeners should show categorical perception of stimuli from the rock-lock continuum in both identification and discrimination, with progressively less categorical perception observed for more experienced, then less experienced Japanese listeners.

In the forced category goodness rating task, native English listeners should rate [w], [j], and the [ɹ] and [l] ends of the continuum as good examples only for the corresponding phonological category (e.g., /w/ for [w]), and as low on each of the other three categories. Ratings for the continuum steps should vary in a similar way to the identification task. The more and less experienced Japanese participants should both rate [w] and [j] as good only for the corresponding phonological category, and low on each of the other three categories. If participants perceive phonological overlap for stimuli along the rock-lock continuum, then forced goodness ratings will be above the lowest rating for more than one phonological category. It is anticipated that the ratings as /r/ and /l/ will be high for both continuum end points, but that the difference between /r/ and /l/ goodness ratings for the same stimulus (e.g., [ɹ]) will be greater for the more than less experienced groups. That is, the overlap should be smaller for more experienced than less experienced Japanese learners of English.

## 2. Materials and Methods

### 2.1. Participants

Japanese native speakers were recruited from Sydney, Australia, via word-of-mouth, noticeboard advertisements at local universities, and a Japanese-language electronic bulletin board service targeted at expatriate Japanese people living in Sydney. The aim was to recruit two samples of Japanese native speakers: (1) migrants who had been living in Australia for a long period, and; (2) recent arrivals. Fifty-five Japanese native speakers were tested. Data were discarded for four participants who were immersed in an English-speaking country at the age 16 or younger, and for one participant who was raised in Hong Kong. To establish a clear difference in length of residence (LOR) between the more experienced and less experienced groups, data were retained for participants who had been living in Australia for a minimum of 2 years, or for 3 months or less. Data for 13 participants with LORs ranging from 5 to 19 months were therefore discarded ($M = 0.89$ years).

The final sample of Japanese native speakers consisted of 19 more experienced English users (16 females, $M_{age}$ = 38 years, Age Range: 21 to 59 years, $M_{LOR}$ = 8 years, LOR Range: 2 to 27 years) and 18 less experienced English users (13 females, $M_{age}$ = 25 years, Age Range: 20 to 35 years, $M_{LOR}$ = 8 weeks, LOR Range: 1 to 13 weeks). The participants were given a small payment for their participation in the study.

The participants were asked to report any languages that they learned outside the home (i.e., in a formal education context) and the age at which they began to learn them. For the more experienced group, all participants began to learn English in Japan between 9 and 13 years of age ($M = 12$ years, $SD = 0.91$ years). One participant did not report the number of years of English study, but the remainder reported between 6 and 15 years ($M = 9$ years, $SD = 2.41$ years). For the less experienced group, the participants began to learn English in Japan between 5 and 13 years of age ($M = 11$ years, SD = 1.9 years). Two participants did not report the number of years of English study, with the remainder completing between 6 and 15 years ($M = 10$ years, $SD = 2.42$ years).

The Australian-English native speakers were recruited from the graduate and undergraduate student population at Western Sydney University, Australia, who received course credit for participation. There were 16 participants (14 females, $M_{age}$ = 21 years, Age Range: 18 to 35 years). Data for an additional four participants were collected but discarded due to childhood acquisition of a language other than English ($n = 1$), self-reported history of a language disorder ($n = 2$), or brain injury ($n = 1$).

## 2.2. Stimuli and Apparatus

Participants were presented with the 10-step /rak/–/lak/ (*rock-lock*) continuum that was first used in MacKain et al. (1981), and the endpoints of the /wak/–/jak/ (*wock-yock*) continuum from Best and Strange (1992); see also (Hallé et al. 1999). The first consonant and vowel portions of the stimuli were generated with the OVE-IIIc cascade formant synthesizer at Haskins Laboratories, and the /k/ was appended to the synthesized syllables. See the original articles for additional stimulus details, including F1, F2, and F3 parameters. A questionnaire was used to collect basic demographic information, and information about the participants' language learning history. The experiment was run on a MacBook laptop running Psyscope X B50 (http://psy.ck.sissa.it/). Participants listened to stimuli through Koss UR-20 headphones set at a comfortable listening level.

## 2.3. Procedure

Participants completed the forced category goodness rating task, followed by /r/–/l/ identification, and AXB discrimination. The language background questionnaire was completed at the end of the session.

*Forced category goodness rating*. Participants were instructed that on each trial they would hear a syllable in their headphones and that their task would be to decide how similar the first sound of the syllable was to one of four English consonant categories, presented on screen (R as in ROCK, L as in LOCK, W as in WOCK, or Y as in YOCK). They were asked to rate the similarity on a 7-point scale, using the numbers on the computer keyboard, where 1 indicated that it was highly similar to the given category, 4 was somewhat similar, and 7 was highly dissimilar.[3] Participants were encouraged to try to use the entire scale from 1–7 across the experiment, and they were instructed not to reflect too long on their response. If they did not respond within 4 s, the trial was aborted, and a message instructed them to respond more quickly. Missed trials were repeated later in the list to ensure that each participant provided a full set of rating data. Each of the 12 stimulus tokens (10 /r/–/l/ continuum steps plus the /w/–/j/ endpoints) were presented five times, each in the context of the four rating categories, resulting in a total of 240 trials. Stimuli were randomized within each of the five blocks. To maintain participant vigilance and to give an opportunity to take a short break, the participant was asked to press the space bar to continue every 20 trials. The task took approximately 20 min to complete.

*Identification*. The 10 steps of the /rak/–/lak/ continuum were used in the identification task. Participants were instructed to listen to syllables through headphones and indicate whether the first sound was more like "r" as in "rock" or "l" as in "lock". They responded using the D and L keys on the keyboard, which were labeled with "R" and "L", respectively. The letters R and L were also displayed on the left- and right-hand side of the screen. Participants were instructed not to reflect for too long on their response. The trial timed out after 2 s, which was followed by a message on screen to respond more quickly. Missed trials were reinserted at a random point in the remaining trial sequence. Participants were presented with 20 randomly ordered blocks of the 10 steps (200 trials in total) and they pressed the space bar at the end of each block to continue. The test took approximately 10 min to complete.

*AXB Discrimination*. Following Best and Strange (1992), participants were tested using AXB discrimination. Three tokens were presented sequentially. The first and third were different steps on the continuum and the middle item was identical to the first or the third token. Participants were tested on steps that differed by three along the continuum ([ɹ]-4, 2-5, 3-6, 4-7, 5-8, 6-9, 7-[l]). Stimuli were presented in all four possible AXB trial combinations (AAB, ABB, BAA, BBA). These 28 trials (7 step pairs × 4 AXB trial combinations) were randomized within blocks, which were presented 5 times. Each step pair was therefore presented 20 times in total across 140 trials. Participants were told that on each trial they

---

3 Although participants had no difficulty using this scale, the scale is reversed in the results section to assist the reader in interpreting the data patterns. That is, 7 is reported as *highly similar* and 1 is reported as *highly dissimilar*.

would hear three syllables in their headphones one after the other, that the first and third syllables were different, and the second syllable was either the same as the first or the third syllable. They were instructed to indicate whether the second syllable was the same as the first or the third syllable, using the keys 1 and 3 on the computer keyboard, basing their decision only on the first sound in each syllable (i.e., the consonant). The software would not allow the participants to respond until the third syllable had finished playing. Participants were instructed not to reflect for too long on their response. A trial timed out after 2 s and was repeated later in the experiment. It took around 5 min to complete.

## 3. Results

### 3.1. Forced Category Goodness Rating

Participants rated the 10 steps of the /r/–/l/ continuum and the endpoints of the /w/–/j/ continuum against the categories /l/, /r/, /w/, and /j/. The results for the three groups are presented in Figure 1. The top-left panel (a) shows the data for less experienced Japanese listeners, the top-right panel (b) for more experienced Japanese listeners, and the bottom panel (c) for Australian-English listeners. The auditory stimulus is plotted on the *x*-axis. Step 1 of the /r/–/l/ continuum is denoted as [ɹ], step 10 is [l], and the other steps are denoted by their step number. The endpoints of the /w/–/j/ continuum are denoted as [w] and [j]. The mean ratings are plotted on the *y*-axis. Thus, the top left point on the English native listener plot represents participants' mean ratings for how well auditory [ɹ] fit the /r/ category, "R"; the bottom left point represents their ratings for auditory [ɹ] to the /l/ category, "L".
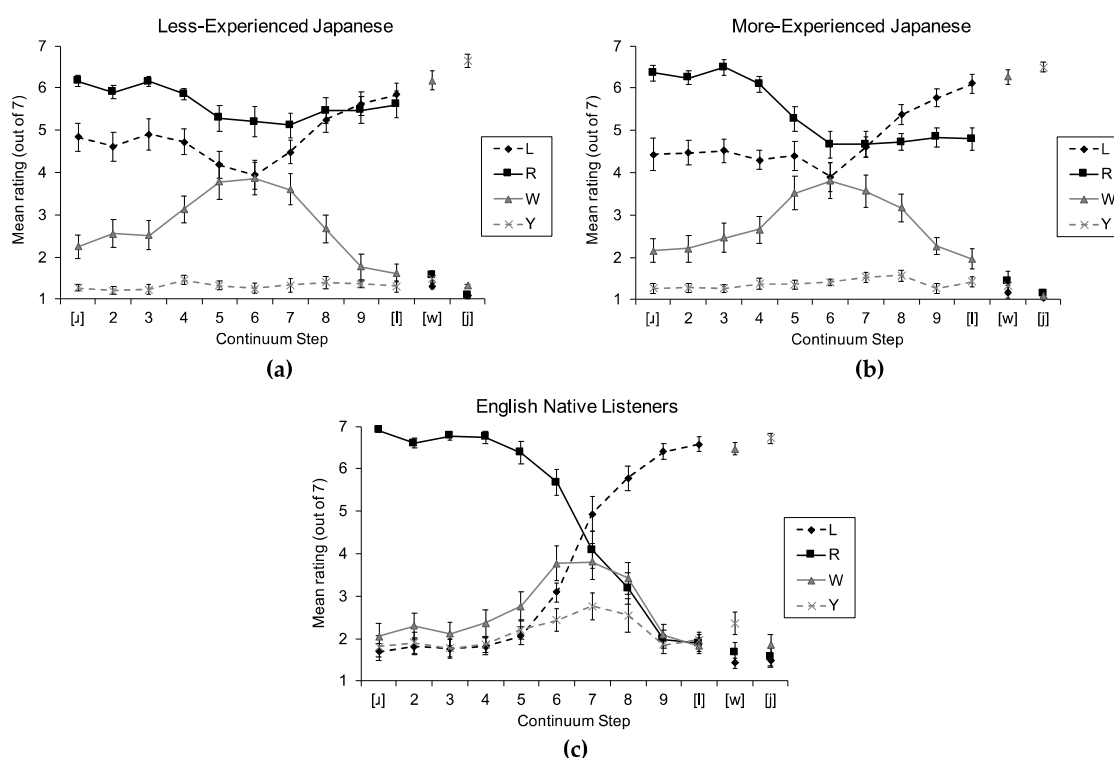


**Figure 1.** Mean forced goodness ratings for each auditory stimulus, to /l/ (L), /r/ (R), /w/ (W), and /j/ (Y) categories, where 1 is "highly dissimilar" and 7 "highly similar". (**a**). Less experienced Japanese (<3 months immersion); (**b**) More experienced Japanese (>2 years immersion); (**c**) Native English listeners. Error bars represent standard error of the mean.

The Australian-English group results, shown in the bottom panel of Figure 1, show a classic categorical perception pattern for the /r/–/l/ continuum for ratings as /l/ and /r/, with the cross-over point between steps 6 and 7. The phones [w] and [j] were each rated as highly similar to /w/ and /j/, respectively, and highly dissimilar to any other category. It

is interesting to note that the ambiguous regions of the /r/–/l/ continuum also appear to have attracted higher acceptability ratings for /w/ and /j/ than the /r/–/l/ endpoints did. To test this, the ratings to /w/ and /j/ for step 7 were compared to those for [l] and [ɹ], respectively, in separate $2 \times 2$ repeated measures analyses of variance (ANOVAs). For step 7 versus [l], there was a main effect of step, $F(1, 15) = 20.79$, $p < 0.001$, $\eta^2_G = 0.27$,[4] and a significant two-way interaction between step and category, $F(1, 15) = 9.09$, $p < 0.001$, $\eta^2_G = 0.06$, indicating that the rating difference between step 7 and [l] was greater for /w/ than for /j/. Simple effects tests showed that the rating difference between step 7 and [l] was significant for both /w/, $F(1, 15) = 20.92$, $p < 0.001$, $\eta^2_G = 0.38$, and /j/, $F(1, 15) = 8.46$, $p = 0.01$, $\eta^2_G = 0.13$. For step 7 versus [ɹ], there were main effects of step, $F(1, 15) = 26.92$, $p < 0.001$, $\eta^2_G = 0.22$, and category, $F(1, 15) = 10.36$, $p = 0.006$, $\eta^2_G = 0.06$, but no interaction, $F(1, 15) = 3.86$, $p = 0.07$. This suggests that both step 7 and [ɹ] are more /w/-like than /j/-like, and that, overall, the ratings for step 7 are higher than those for [ɹ]. Together, these results indicate that the most ambiguous step was perceived as somewhat /w/-like and weakly /j/-like, in addition to being perceived as somewhat /r/- and /l/-like. The [l] endpoint was not perceived as similar to /w/ or /j/, and it appears that the listeners may have perceived the [ɹ] endpoint to have some similarity to /w/ but not /j/.

An initial comparison of the pattern of results for the two Japanese groups, in the top-left and top-right panels of Figure 1, suggests that there may be a difference in their ratings of the /r/-/l/ continuum. Considering first the endpoints, the less experienced group appear to rate [l] as both an acceptable /l/ and an acceptable /r/, whereas [ɹ] appears to be rated as a more acceptable /r/ than /l/. This supports the idea that they perceive a phonological overlap between /r/ and /l/ for both [ɹ] and [l], but at first glance it suggests that they already have a reasonable sensitivity to the difference between /r/ and /l/ at the [ɹ] end of the continuum. It is important to note, however, that the ratings as /r/ across the entire continuum are uniformly high. The separation in ratings for [ɹ] as /r/ and as /l/ may be due to that phone being perceived as a poorer /l/ rather than a more acceptable /r/. The more experienced Japanese group, on the other hand, appear to have rated [ɹ] as a more acceptable /r/ than /l/, and [l] as a more acceptable /l/ than /r/. The lower of the two goodness ratings for both [ɹ] and [l] are around 4—"Somewhat Acceptable". There is also a remarkable similarity between the shape of the response curve for ratings as /l/ for both groups at the /l/-end of the continuum, which is consistent with the idea that they have established a common L1-L2 category for English /l/ and Japanese /ɾ/.

To test these observations, the ratings as /r/ and /l/ for the 10 steps of the /r/–/l/ continuum were subjected to a $2 \times (2) \times (10)$ mixed ANOVA. The between-subjects variable was *group* (less experienced vs. more experienced Japanese listeners) and the two within-subjects variables were *category* (/l/ vs. /r/) and *step*. There were main effects of category, $F(1, 35) = 40.47$, $p < 0.001$, $\eta^2_G = 0.09$, and step, $F(9, 315) = 12.17$, $p < 0.001$, $\eta^2_G = 0.09$, and a significant two-way interaction between them, $F(9, 315) = 18.96$, $p < 0.001$, $\eta^2_G = 0.13$. The differential responding by the more experienced and less experienced groups, that can be seen in Figure 1, was confirmed by a significant three-way interaction between category, step, and group, $F(9, 315) = 2.67$, $p = 0.005$, $\eta^2_G = 0.02$. To explore the three-way interaction further, separate two-way ANOVAs were run for each group. There were two-way interactions between category and step for both the more experienced, $F(9, 162) = 14.93$, $p < 0.001$, $\eta^2_G = 0.19$, and the less experienced groups, $F(9, 153) = 5.20$, $p < 0.001$, $\eta^2_G = 0.07$. Another set of two-way ANOVAs was conducted to test whether the groups differed for each category. There was a two-way interaction between step and group for /r/, $F(9, 315) = 2.59$, $p = 0.007$, $\eta^2_G = 0.04$, but not for /l/, $F(9, 315) = 0.68$, $p = 0.73$, suggesting that the differences between the groups can be accounted for by improvements in perception of their /r/ category only. To test for differences in goodness ratings as /r/

---

4    Generalized eta squared ($\eta^2_G$) is a measure of effect size that is appropriate for mixed designs (Olejnik and Algina 2003). It is compatible with Cohen's (1988) benchmarks for interpreting eta squared (small = 0.01, medium = 0.06, and large = 0.14).

versus /l/ at each continuum step, post-hoc paired *t*-tests were run separately for each group, with a Bonferroni-adjusted alpha level of 0.005. The results are presented in Table 1. For the less experienced group, there are significant differences between ratings as R versus L at steps [ɹ] through 4, and for step 6. For the more experienced group, the ratings are also different for steps [ɹ] through 4. Importantly, and unlike the less experienced group, they are also different for step [l].

**Table 1.** Post-hoc paired *t*-tests for ratings as /r/ versus /l/ at each continuum step, for less experienced and more experienced Japanese listeners.

| Group | | Step | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | [ɹ] | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | [l] |
| Less experienced | *t*(17) | −3.43 | −3.84 | −3.65 | −4.30 | −2.66 | −4.81 | −2.84 | −1.37 | 0.75 | 0.71 |
| | *p* | **0.003** | **0.001** | **0.002** | **<0.001** | 0.016 | **<0.001** | 0.011 | 0.190 | 0.462 | 0.488 |
| More experienced | *t*(18) | −6.22 | −4.42 | −5.58 | −5.02 | −2.53 | −2.40 | −0.19 | 2.07 | 2.34 | 4.11 |
| | *p* | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.021 | 0.028 | 0.850 | 0.054 | 0.031 | **0.001** |

Values in boldface are significant at a Bonferroni-adjusted alpha rate of 0.005.

Like the Australian-English listeners, both Japanese groups rated [w] and [j] as highly similar to /w/ and /j/, respectively, and not to any other category. The ratings as /w/ across the /r/–/l/ continuum appear to be similar in the groups. They also appear to have similar ratings as /j/, but with a flatter response than the English native listeners. The ratings as /w/ and /j/ for step 7 and [l], and for step 7 and [ɹ], were compared for the two Japanese groups using separate $2 \times (2) \times (2)$ ANOVAs. The between-subjects variable was *group* (less experienced vs. more experienced) and the within-subjects variables were *category* (/w/ vs. /j/) and *step* (step 7 vs. [l] or step 7 vs. [ɹ]). For step 7 versus [l], there were main effects of category, $F(1, 35) = 35.77$, $p < 0.001$, $\eta^2_G = 0.20$, and step, $F(1, 35) = 21.22$, $p < 0.001$, $\eta^2_G = 0.10$, and a significant two-way interaction between them, $F(1, 35) = 15.90$, $p < 0.001$, $\eta^2_G = 0.07$. Crucially, there was no three-way interaction between category, step, and group, $F(1,35) = 0.05$, $p = 0.83$, suggesting that the more and less experienced Japanese listeners responded similarly to each other. The significant two-way interaction was further probed with tests of simple effects, which showed that the differences in ratings for step 7 and [l] were significant for ratings as /w/, $F(1, 35) = 22.08$, $p < 0.001$, $\eta^2_G = 0.03$, but not for ratings as /j/, $F(1, 35) = 1.37$, $p = 0.25$. The results were similar for step 7 versus [ɹ], with the main effects of category, $F(1, 35) = 48.30$, $p < 0.001$, $\eta^2_G = 0.27$, and step, $F(1, 35) = 13.73$, $p = 0.001$, $\eta^2_G = 0.06$, and a significant two-way interaction between them, $F(1, 35) = 4.18$, $p < 0.001$, $\eta^2_G = 0.02$.

*3.2. Identification*

The mean percent L responses for identification of steps from the /r/–/l/ continuum are presented in Figure 2. The native English listeners show the classic ogive-shaped categorical perception function that was observed in previous studies using the same stimuli (Best and Strange 1992; Hallé et al. 1999; MacKain et al. 1981). The boundary location appears to be between steps 6 and 7, mirroring the results from the forced category goodness rating task. This seems to be closer to the /l/ end of the continuum than the American-English listeners in the previous studies, whose boundary location was between steps 5 and 6. The more experienced Japanese participants appear to have a shallower function than the native English participants and the endpoints of the continuum do not reach 0% or 100% (at 19% and 77%, respectively). The less experienced Japanese participants' function appears to be even shallower than the more experienced participants' and their endpoints are closer to the chance level of 50% ([ɹ] = 32%, [l] = 58%).
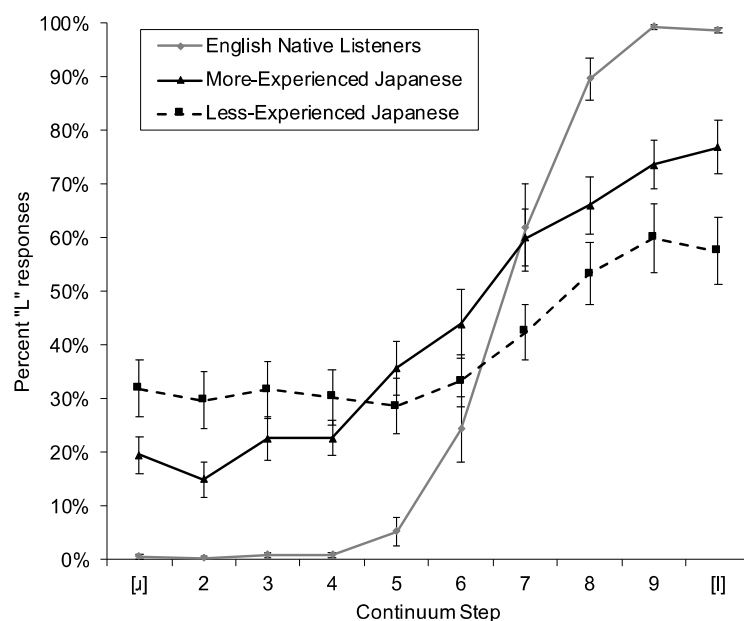
**Figure 2.** Mean percent "L" responses from the R-L identification task for the three participant groups. Error bars represent standard errors of the mean.

Given the English listeners' ceiling performance for steps [ɹ]-4, 9, and [l], and the corresponding lack of variability, it is not appropriate to use standard parametric tests to compare their performance to the Japanese groups'. It is clear that they perform differently from the more experienced Japanese listeners. Furthermore, given that the shapes of the response curves for most of the Japanese participants in this data set do not appear to follow an ogive-shaped function, the data have not been fit to a cumulative distribution function, as they were in previous studies (Best and Strange 1992; Hallé et al. 1999). A 2 × (10) ANOVA was conducted to test whether the response curves differed for more experienced versus less experienced Japanese listeners. The between-subjects factor was *experience* and the within-subjects factor was *step*. The shape of the response curve was tested using orthogonal polynomial trend contrasts on the step factor—linear, quadratic (one turning point), cubic (two turning points), and quartic (three turning points) (Winer et al. 1991, for contrast coefficients). There was no significant main effect of experience, $F(1, 35) = 0.75$, but there were significant overall linear, $F(1, 35) = 60.39$, $p < 0.001$, quadratic, $F(1, 35) = 7.89$, $p = 0.008$, cubic, $F(1, 35) = 17.76$, $p < 0.001$, and quartic trends, $F(1, 35) = 5.34$, $p = 0.02$. Those significant trend contrasts simply indicate that the curve has a complex shape. The important question is whether there are any significant interactions between the trend contrasts and experience. The only significant interaction was with the linear trend contrast, $F(1, 35) = 7.91$, $p = 0.008$. Therefore, while there is no evidence for a difference in the shape of the Japanese participants' response curve, the significant interaction shows that the more experienced group have a generally steeper function than the less experienced group, as can be seen in Figure 2.

### 3.3. Discrimination

Mean percent correct responses for discrimination of the seven pairs of steps from the continuum are presented in Figure 3. Again, there are clear differences between the English and Japanese listeners' performance. The English listeners show the classic categorical perception discrimination response, with poorer discrimination for steps that are on the same side of the categorical boundary and more accurate discrimination for steps that cross the category boundary. The Japanese participants, on the other hand, show a clear double peak.
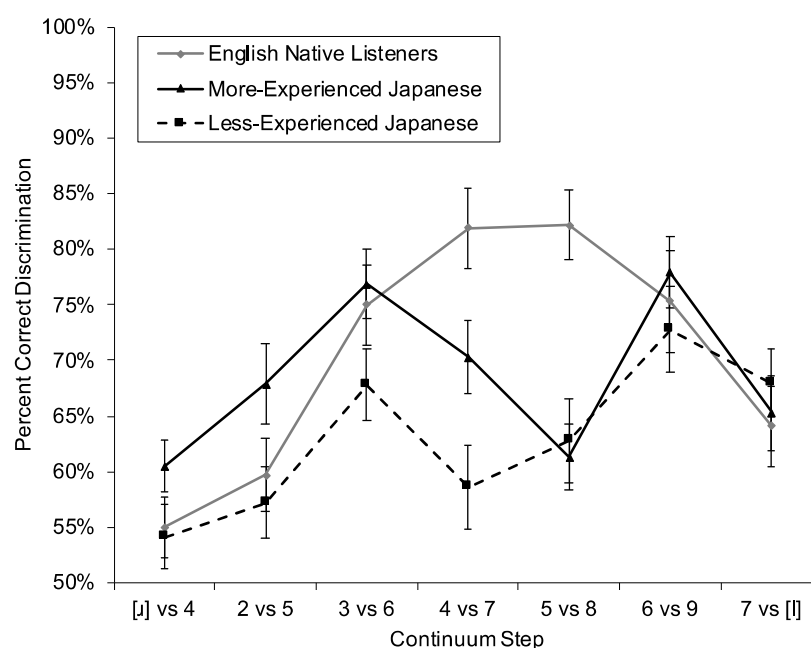
**Figure 3.** Mean percent correct discrimination of pairs of steps from the /r/–/l/ continuum for the three participant groups. Error bars represent standard errors of the mean.

As the performance of the native speakers is not at ceiling on this task, it is possible to conduct a 3 × (7) ANOVA comparing all three groups. The between-subjects factor was *group*, with two planned contrasts. The *language background* contrast compared the English listeners with the combined results of the two Japanese groups and the *experience* contrast compared the two Japanese groups only. The within-subjects factor was *step pair*, and the shape of the curve was tested again using orthogonal polynomial contrasts. The language background contrast was not significant, $F(1, 50) = 3.47$, but a significant experience contrast showed that the more experienced group performed more accurately, overall, than the less experienced group, $M_{\text{difference}} = 5.53\%$, $SE = 2.74$, $F(1, 50) = 4.08$, $p = 0.049$. There was a significant overall linear trend, $F(1, 50) = 27.48$, $p < 0.001$, reflecting the gradual increase in accuracy, collapsed across all three groups, from the [ɹ]-end to the [l]-end of the continuum. There was no interaction between the linear trend contrast and language background, but it interacted significantly with experience, $F(1, 50) = 4.49$, $p = 0.04$. This can be seen in Figure 3, where the less experienced group appear to be relatively less accurate at the [ɹ]-end than the [l]-end of the continuum, whereas the more experienced group show more similar levels of accuracy at both ends. There was also a significant overall quadratic trend, $F(1, 50) = 33.81$, $p < 0.001$, which interacted with the language background contrast only, $F(1, 50) = 20.90$, $p < 0.001$. This reflects the fact that the English participants' responses show a quadratic shape, whereas the Japanese participants' responses appear to follow a quartic shape. Indeed, there was no significant cubic trend, and no interactions, and no significant overall quartic trend, but there was a significant interaction between the quartic trend and language background, $F(1, 50) = 9.67$, $p = 0.003$.

## 4. Discussion

The aim of this study was to test whether certain L2 learners, who initially acquired their L2 in a formal learning context, might perceive an L2 phone as an instance of more than one L2 phonological category, and whether that perceived phonological overlap is smaller with longer versus shorter periods of L2 immersion. To test for perceived phonological overlap, participants completed a forced category goodness rating task, where they were presented with an auditory token and rated its goodness of fit to a given English phonological category label, L, R, W, or Y. It was hypothesized that Japanese native speakers who had first been exposed to English in Japan would perceive a phonological

overlap between /r/ and /l/, but that the overlap would be smaller for those with a long period of immersion in an English-speaking environment (>2 years) than those with a short period of immersion (<3 months). Native English speakers rated [ɹ] as a highly similar to /r/, but not /l/, /w/, or /j/, and [l] as highly similar to /l/, but not to the other three categories. In contrast, the Japanese native speakers rated [ɹ] as highly similar to /r/, moderately similar to /l/, and dissimilar to /w/ and /j/. The two groups of Japanese speakers differed from each other in their ratings of [l]. The less experienced Japanese group rated [l] as highly similar to *both* /r/ and /l/, whereas the more experienced Japanese group rated [l] as highly similar to /l/ and only moderately similar to /r/. In short, the hypotheses were supported. When the Japanese listeners perceived stimuli along the continuum between /r/ and /l/, they perceived varying degrees of phonological overlap with /r/ and /l/, and the overlap was smaller for the more experienced than the less experienced group. This study is cross-sectional, so it is not possible to conclude definitively that there is a reduction in overlap due to immersion experience, but these results are certainly consistent with that idea.

The general pattern of results observed for the less experienced group was predicted by a learning scenario, proposed here, that had not previously been considered by Best and Tyler (2007) when they outlined PAM-L2. The scenarios presented in Best and Tyler were based on functional monolinguals who were acquiring an L2 in an immersion setting. In contrast, the participants in both Japanese groups had learned English in Japan prior to immersion. As the /r/–/l/ contrast is a category-goodness assimilation for Japanese native speakers (e.g., Guion et al. 2000), the PAM-L2 learning scenario suggests that [l] would initially be perceived as a common L1-L2 phonetic category with Japanese [ɾ], and that Japanese /ɾ/ and English /l/ would form a common L1-L2 phonological category. English [ɹ] would first be perceived as an allophonic variant of the common L1-L2 /ɾ/–/l/ category, and then a new L2-only /r/ phonological category would be established when the learner recognized that the phonetic contrast between /r/–/l/ signaled a phonological distinction. Thus, the non-native category-goodness contrast would become an L2 two-category contrast. Such a situation should have resulted in a data pattern that resembled the native English speakers' results. Since the Japanese native speakers in this study acquired English in a formal learning situation in Japan, it was argued here that they may have needed to establish a new phonological category for /r/ before they had managed to tune in to its phonetic properties. That learning scenario is consistent with the data pattern observed in this study. The participants were able to indicate that steps along the continuum were perceived as similar to /r/, but they rated the same stimuli as also having various degrees of similarity to /l/. The fact that they rated the steps as being dissimilar to /j/ shows that they were not simply indicating that any L2 consonant was similar to /r/. Rather, it is consistent with the idea that they had developed a new phonological category for /r/ that was poorly tuned to the phonetic properties that distinguish it from /l/. In PAM terms, both /r/ and /l/ would be uncategorized clustered assimilations. Rather than becoming a two-category assimilation, /r/–/l/ would be a completely overlapping uncategorized-uncategorized contrast (Faris et al. 2016, 2018).

A close inspection of Figure 1 shows that the response pattern for ratings as /l/ appears to be similar for the less and more experienced groups, with stimuli closer to [ɹ] rated as poorer /l/s than those closer to the [l] end. Indeed, follow-up analyses of the three-way interaction between group, category, and step showed that the difference between the two groups was entirely due to differences in ratings as /r/. The less experienced group gave high ratings as /r/ across the continuum, whereas the more experienced group gave lower ratings as /r/ towards the [l]-end of the continuum. The finding that the ratings as /l/ were unaffected by immersion experience fits with the idea that English /l/ is a common L1-L2 category with Japanese /ɾ/, as an L1 category may be more resistant to change with L2 experience than a new L2-only phonological category. Future research could investigate this further by including Japanese /ɾ/ as a rating category and auditory stimulus, in addition to English /r/ and /l/.

Presenting the /r/–/l/ continuum, rather than isolated tokens of /r/ and /l/, allows us to compare the internal structure of the phonological categories. It is clear from comparing the three groups in Figure 1 that the English native listeners have a much clearer separation of their /r/ and /l/ categories than the Japanese listeners. Steps [ɹ]-4 have phonetic properties that appear to be prototypically /r/ for the English listeners, and those of steps 9-[l] appear to be prototypically /l/. The uniformly high ratings as /r/ across the continuum for the less experienced Japanese group suggests that their L2 English /r/ category may initially cover a broad region of phonetic space, with minimal differences in phonetic goodness of fit. Improvements that have been observed previously after periods of immersion may, therefore, be due more to improvements in category tuning to /r/ than /l/ (e.g., MacKain et al. 1981). Training regimes for improving /r/–/l/ perception in Japanese native listeners have traditionally focused on identifying minimal-pair words (e.g., Bradlow et al. 1997; Hattori and Iverson 2009; Lively et al. 1993; Lively et al. 1992; Lively et al. 1994), that is, on recognizing those phonetic characteristics that define category membership of /r/ versus those that define /l/. If the learner perceives a phonological overlap, and the token is perceived as an equally good example of both /r/ and /l/, then the utility of that training may be limited. The findings of the present study suggest that there may be some benefit in training learners to recognize which tokens should be a good versus poor fit to a category rather than identifying which category they belong to, before transitioning to training regimes that focus on detecting the relational phonetic differences that characterize phonological distinctions (e.g., using a discrimination task).

### 4.1. Identification and Discrimination

In contrast to the forced category goodness rating task, where participants rated the phonetic goodness of fit to a phonological category that was provided on each trial, the identification task required them to make a forced choice between two phonological categories. In an identification ask, a participant could decide that a given stimulus was /r/, for example, because either it clearly sounded like /r/ or it clearly did *not* sound like /l/. As the forced category goodness rating task provides some insight into the listeners' judgement of the extent to which each continuum step resembled or did not resemble /l/ or /r/, their ratings should be related to their identification response function. As the forced category goodness ratings as /l/ and /r/ are basically mirror images of each other for the English native listeners, they could have made their decision using either criterion, and their identification response curve would have been similar to the one presented in Figure 2. For both Japanese groups, the separation in forced category goodness ratings as /l/ versus /r/ is wider at the [ɹ]-end than the [l]-end of the continuum. This is reflected in the identification accuracy, given that both groups were more accurate at identifying [ɹ] than [l] (see Figure 2). That result is consistent with Hattori and Iverson (2009), and with a recent study where Japanese listeners were more accurate for /r/ than /l/ in a *ranby* versus *lanby* identification task (Kato and Baese-Berk 2020). The authors of that study suggested that the asymmetry was due to a bias towards the category that was more dissimilar to the closest native category. The results of this study complement that conclusion by providing a novel theoretical explanation for the source of the bias. Japanese listeners identify /r/ more accurately than /l/ because the perceived phonological overlap between /r/ and /l/ is less pronounced for /r/ than it is for /l/. It would be interesting to see whether the same pattern of overlap is observed for the more dissimilar phone in category-goodness contrasts from other languages and language groups.

Given that the two groups did not differ on their ratings as /l/ in the forced category goodness rating task, they should have had similar judgements in the identification task about what did and did not sound like /l/. Relative differences between the groups' identification function should be attributable to their divergent ratings of each step as /r/. Indeed, all of the steps were acceptable-sounding versions of /r/ for the less experienced group (Figure 1a) and the shape of their identification function (Figure 2) is strikingly similar to that for their ratings as /l/. The more experienced group appears to have a

slightly wider separation than the less experienced group between ratings as /r/ and /l/ for steps [ɹ]-4 (Figure 1b), which is reflected in identification by a lower percentage of "L" responses on steps [ɹ]-4. Ratings as /r/ dropped and remained steady from steps 5-[l] (Figure 1b) and this corresponds to the relatively higher accuracy of the more versus less experienced group on identification of those steps (Figure 2). Thus, the forced category goodness rating task has provided insights into the reason why Japanese listeners do not show categorical perception across an /r/-/l/continuum. Differences in goodness of fit to simultaneously perceived phonological categories modulates their judgements of whether the stimulus is or is not a member of each category.

MacKain et al. (1981) compared identification and discrimination of the same continuum by Japanese native speakers living in the US who had undergone intensive conversation training in English, another group with little or no such training, and native speakers of English. The experienced group in that study showed categorical perception along the rock-lock continuum, with an identification response that did not differ from the native speakers', whereas the inexperienced group showed a fairly flat response function. The identification results in the present study were comparable to MacKain et al. for the less experienced group, but the more experienced group was not similar to the native speakers in this study. One key difference between the studies was the criterion used to select the more experienced group. Whereas MacKain et al. selected participants on the basis of conversational training, here they were selected on the basis of length of residence. Focused conversational training in an immersion situation may have resulted in more native-like L2-learning outcomes than simply residing in an L2 environment. A comparison of the present results with those of MacKain et al. would support Flege's (2009, 2019) contention that access to quality native-speaker input is a more direct predictor of L2 perceptual learning than length of residence.

In discrimination, the English speakers showed a clear peak across the categorical boundary, in line with previous research. There is a clear double peak for both Japanese groups, which is in contrast to the relatively flat distribution observed in MacKain et al. (1981). The identification results do not seem to provide any explanation for why a double peak was observed. For example, the less experienced group showed a peak for step 3 versus step 6, but these were identified similarly. A double peak is an indication that the participants may have perceived a third category in the middle of the continuum. The forced category goodness ratings suggest that there was some degree of /w/ perceived in the middle of the continuum. Although it may seem unlikely that participants would have identified /w/ in the middle of the continuum when the goodness ratings for /w/ were no higher than they were for /r/ or /l/, previous findings of /w/ identification in the middle of other /r/−/l/ continua (Iverson et al. 2003; Mochizuki 1981) make this a plausible explanation. Another possibility is that they perceived a different category in the middle of the continuum (e.g., their native Japanese /ɾ/ or possibly the vowel-consonant sequence /ɯɾ/, see Guion et al. 2000), but as they were not asked to rate the stimuli against other categories, that is a question that would need to be addressed in future research. Discrimination accuracy generally increased for the less experienced group along the continuum from [ɹ] to [l], whereas the relatively more accurate discrimination of the more experienced group was fairly level. This may reflect the more experienced group's greater sensitivity to goodness differences in both /r/ and /l/, whereas the less experienced group may have relied primarily on goodness differences relative to /l/ only.

### 4.2. Conflicting Findings between Pre-Lexical and Lexical Tasks

As phonological categories are pre-lexical perceptual units for PAM-L2, perceived phonological overlap would be a logical consequence of acquiring a phonological category before sufficient perceptual learning had taken place to differentiate it from other categories in the phonological system. The results of this study, and of Kato and Baese-Berk (2020), are consistent with that account. However, in a study examining the time course of L2 spoken word recognition, Cutler et al. (2006) observed an asymmetry that appears to be the

reverse of the one observed here in pre-lexical tasks. Japanese and English native speakers heard an instruction to click on one of four objects presented on the screen while an eye tracker monitored their eye movements. On critical trials, one picture depicted an object containing /r/ (e.g., *writer*), another containing /l/ (e.g., *lighthouse*), and there were two non-competitor pictures containing neither /r/ nor /l/. The /r/-/l/ word pairs were chosen so that the onsets overlapped phonologically, such that the participants would need wait for disambiguating information (e.g., the /h/ of *lighthouse*) if they were unable to tell /r/ and /l/ apart. When the word containing /r/ was the target, the Japanese participants took longer to settle their gaze on the correct object than the English native speakers did, suggesting that they were unable to disambiguate the words on the basis of /r/ or /l/. However, when the /l/-word was the target, they settled on the correct picture early, at the same point in time as the English native speakers. Thus, there was an asymmetry in word recognition, such that they were apparently more efficient at recognizing words beginning with /l/ than those beginning with /r/ (see Weber and Cutler 2004, for similar results on Dutch listeners' recognition of English words containing /ɛ/ and /æ/). Given that perceived phonological overlap in this study was smaller for /r/ than /l/, and /r/ is identified more accurately than /l/ (Kato and Baese-Berk 2020), it is surprising that spoken word recognition should show an asymmetry in the opposite direction (see Amengual 2016; Darcy et al. 2013 for other examples of a mismatch between performance pre-lexical and lexical tasks). Cutler et al. explained their results in terms of lexical processing, rather than perception of phonological categories, which may account for the difference. They suggested that the Japanese listeners had established lexical entries that preserved the /r/–/l/ phonological distinction, even though they could not reliably discriminate the contrast, and provided two possible explanations for the asymmetry. One possibility (also suggested by Weber and Cutler 2004) is that when /r/ is included in a lexical entry, it does not receive any bottom-up activation from speech, and nor does it inhibit (or is it inhibited by) the activation of other words as they compete for selection as the most likely word candidate. Activation of the word containing /r/ (and inhibition of other competitors) would only proceed via input that matched its other phonemes. The second possibility is that both /l/ and /r/ words contain the L1 /ɾ/ category in their lexical entries. Words containing /l/ would be activated when a reasonable sounding /ɾ/ is encountered and those containing /r/ would be activated when encountering a poorer match. By that account, the asymmetry arose because /r/ would never be perceived as a reasonable match for /ɾ/, but /l/ could be perceived as a poorer match for /r/. Thus, /l/ would only ever contact words in the lexicon containing /l/, but there is a reasonable probability that /r/ would contact words containing both /l/ and /r/.

Darcy et al. (2013) also concluded that lexical encoding was responsible for the asymmetry. They showed that, in spite of accurate discrimination of L2 Japanese singleton-geminate contrasts or German front-back vowel contrasts, lexical decision performance for L2 learners was poorer than it was for native speakers, particularly for nonword items. They also observed an asymmetry in lexical decision. The stimulus words contained either a more or less native-like L2 phoneme and the nonwords were created by swapping the target phoneme with the other member of the pair (e.g., the German word for 'honey', *Honig* /honɪç/, became the nonword *\*Hönig* /hønɪç/). Accuracy for words was higher when the category was more versus less nativelike, and accuracy for minimal-pair nonwords was higher when the category was less versus more native-like. Similar to Cutler et al. (2006), Darcy et al. concluded that lexical coding for the less native-like category is fuzzy, and that it encodes the goodness of fit to the dominant L1 category. Interestingly, advanced German learners did not show the asymmetry, which suggests that lexical encoding can improve with L2 experience.

PAM-L2 (Best and Tyler 2007) may provide a slightly different perspective to the conclusions of Cutler et al. (2006) and Darcy et al. (2013). For PAM (Best 1995) and PAM-L2, phonological categories are perceptual, and they are the result of attunement to the higher-order phonetic properties that are relevant both for recognizing words and

for telling them apart from other words in the language. As /l/ is initially perceived by Japanese native listeners as a good instance of /ɾ/, their existing L1 /ɾ/ category would be used for acquiring any English words containing /l/ (a common L1-L2 category), and a new L2-only phonological category would be established for English /r/. In spoken word recognition, then, words containing /l/ would benefit from an existing L1 category that is already integrated into processes of lexical competition. In contrast, it may take some time for words containing a new L2 category (i.e., /r/) to establish inhibitory connections that would reduce the activation of competitor words (as may have eventually occurred for the advanced German learners in Darcy et al. 2013). Thus, when the Japanese native listeners in Cutler et al. (2006) perceived /l/, their native lexical competition processes would have inhibited activation of competitor words, including the one containing /r/. They may also have perceived /r/ pre-lexically, but without the benefit of inhibitory connections to other words, the /r/ competitor word would have been inhibited by the word containing /l/. For target words containing /r/, both the /r/- and /l/-words would be activated, but the poorer fit to the /l/ category would limit the activation of the /l/-word competitor. Thus, both candidates remained activated until disambiguating information was encountered. Clearly, more research needs to be done to tease apart the pre-lexical and lexical influences on L2 speech perception.

### 4.3. Methodological Considerations

The forced category goodness rating task was devised for this study because categorization with goodness rating might underestimate the perceived phonological overlap. To illustrate, Japanese participants rated step 8 as having various degrees of similarity to /l/, /r/, and /j/, but the ratings for /j/ were lower than for the other two categories. Had they completed a categorization test first, they may not have selected "Y" at all because the other three categories are clearly a better fit. A categorization task was not included for comparison here because the session was already quite long. Nevertheless, it is clear that the forced category goodness rating task is capable of detecting category overlap, and that category overlap was observed for both the native and non-native listeners.

The success of the forced category goodness rating task at detecting perceived phonological overlap raises the question of whether it should be adopted in favor of the standard categorization with goodness rating task. Indeed, Faris et al. (2018) suggested that it might be necessary to reconsider the use of arbitrary categorization criteria and the forced category goodness rating task removes the necessity of specifying a threshold for categorization. It may also solve a problem with the categorization of vowels; some participants have difficulty using the keyword labels for identifying vowels, particularly for a language like English, where some of the grapheme–phoneme correspondences are ambiguous. Faris et al. familiarized participants with the 18 English vowel labels, using English vowel stimuli and providing feedback, but they found that up to a quarter of the participants had difficulty selecting the correct label. While this could mean that those participants had difficulty categorizing native vowels, a more likely possibility is that they had poor phonological awareness, and that affected their ability to perform well on that metalinguistic task. Forced category goodness rating might alleviate that problem because the participants are provided with the category against which to judge the auditory stimuli and they do not need to search for the category that corresponds to the vowel that they heard. However, one clear limitation of the forced category goodness rating is that it is much more labor-intensive than categorization. In the case of Faris et al., participants would need to have rated 32 Danish vowels multiple times against the 18 English vowel categories. This would have resulted in thousands of trials. This is not to say that a forced category goodness rating task should be avoided. If it proved to provide a more accurate estimation of perceptual assimilation, then researchers would need to devote the time necessary to collecting those data.

If the forced category goodness rating task was adopted as a test of perceptual assimilation, then there would no longer be arbitrary thresholds for determining whether

a non-native phone was assimilated as categorized to the native phonological system. Instead, a non-native phone could be deemed to be categorized as a given native phonological category if the mean rating of the stimulus to that category was significantly above the lowest possible rating (e.g., 1 out of 7, where the 1 is defined as "no similarity"). Expanding on Faris et al. (2016), non-native phones would be categorized as focalized if only one category had a non-negligible rating, categorized as clustered if more than one category had a non-negligible rating, or uncategorized as dispersed if no category had a non-negligible rating.

In spite of the potential for this task to support future theoretical advances in research on cross-language and second-language speech perception, careful comparisons need to be made between the data obtained from current methods and the forced category goodness rating task before suggesting a change to standard research protocols. For example, the English native speakers in this study did not give the lowest possible score for [ɹ] as /w/ (2.05 out of 7), even though they clearly perceived that stimulus as belong to their /r/ category ([ɹ] as /r/ was rated at 6.90 out of 7). This suggests that ratings at the lower end of the scale may reflect phonetic similarity rather than phonological category membership. It may nevertheless be necessary to ask participants to make a decision about category membership rather than relying solely on a goodness-of-fit judgement. In fact, Tyler (2021) has identified four different sources of information that non-native listeners could use to discriminate non-native phones. Any new method for assessing perceptual assimilation would need to be capable of assessing listeners' sensitivity to any information available to them for discriminating given non-native contrast. Until careful methodological studies have been completed comparing different approaches to categorization, studies testing PAM/PAM-L2 predictions should continue to use categorization with goodness rating, giving participants the opportunity to select from among all possible vowel or consonant categories (Bundgaard-Nielsen et al. 2011a, 2011b; Faris et al. 2016).

## 5. Conclusions

In an ideal learning situation, adults would tune in to the phonetic and phonological properties of an L2 prior to establishing a large L2 vocabulary. However, this is not the way that L2s are generally learned. Classroom-based learning is more common and it provides opportunities to learn about phonological distinctions before attuning to the phonetic properties that define phonological categories and distinguish them from each other. It was argued here that such a situation may give rise to perceived phonological overlap between L2 categories. The results of a forced category goodness rating task showed that Japanese native speakers who first acquired English in the classroom perceived varying degrees of phonological overlap between English /r/ and /l/ when they encountered either category in speech. The overlap was smaller for those with more than two years of immersion experience, as compared to those with less than three months, suggesting that learners continue to attune to the phonological distinction with appropriate input. Assessment of perceived phonological overlap in L2 learners may help with tracking phonological development and with tailoring perceptual training to those contrasts where lexically guided perceptual retuning is most likely to be effective (see Tyler 2019, for a discussion of how PAM-L2 might apply to classroom foreign language acquisition). Future research should investigate category overlap using natural stimuli, and test whether discrimination accuracy for /r/ and /l/ can be predicted by the degree of phonological overlap between the L2 /r/ and /l/ categories.

**Institutional Review Board Statement:** The protocol was approved by the University of Western Sydney Human Research Ethics Committee (Approval 07/040).

**Informed Consent Statement:** All subjects gave their informed consent for inclusion before they participated in the study.

## References

Amengual, Mark. 2016. The perception of language-specific phonetic categories does not guarantee accurate phonological representations in the lexicon of early bilinguals. *Applied Psycholinguistics* 37: 1221–51. [CrossRef]

Antoniou, Mark, Catherine T. Best, and Michael D. Tyler. 2013. Focusing the lens of language experience: Perception of Ma'di stops by Greek and English bilinguals and monolinguals. *Journal of the Acoustical Society of America* 133: 2397–411. [CrossRef] [PubMed]

Antoniou, Mark, Michael D. Tyler, and Catherine T. Best. 2012. Two ways to listen: Do L2-dominant bilinguals perceive stop voicing according to language mode? *Journal of Phonetics* 40: 582–94. [CrossRef]

Aoyama, Katsura, James E. Flege, Susan G. Guion, Reiko Akahane-Yamada, and Tsuneo Yamada. 2004. Perceived phonetic dissimilarity and L2 speech learning: The case of Japanese /r/ and English /l/ and /r/. *Journal of Phonetics* 32: 233–50. [CrossRef]

Best, Catherine T. 1994a. The emergence of native-language phonological influences in infants: A perceptual assimilation model. In *The Development of Speech Perception: The Transition from Speech Sounds to Spoken Words*. Edited by Judith C. Goodman and Howard C. Nusbaum. Cambridge: MIT Press, pp. 167–244.

Best, Catherine T. 1994b. Learning to perceive the sound pattern of English. In *Advances in Infancy Research*. Edited by C. Rovee-Collier and Lewis P. Lipsitt. Norwood: Ablex, pp. 217–304.

Best, Catherine T. 1995. A direct realist view of cross-language speech perception. In *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. Edited by Winifred Strange. Baltimore: York Press, pp. 171–204.

Best, Catherine T., and Gerald W. McRoberts. 2003. Infant perception of non-native consonant contrasts that adults assimilate in different ways. *Language and Speech* 46: 183–216. [CrossRef]

Best, Catherine T., Gerald W. McRoberts, and Elizabeth Goodell. 2001. Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *Journal of the Acoustical Society of America* 109: 775–94. [CrossRef]

Best, Catherine T., Gerald W. McRoberts, and Nomathemba M. Sithole. 1988. Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance* 14: 345–60. [CrossRef]

Best, Catherine T., and Winifred Strange. 1992. Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of Phonetics* 20: 305–30. [CrossRef]

Best, Catherine T., and Michael D. Tyler. 2007. Nonnative and second-language speech perception: Commonalities and complementarities. In *Second Language Speech Learning: The Role of Language Experience in Speech Perception and Production*. Edited by Murray J. Munro and Ocke-Schwen Bohn. Amsterdam: John Benjamins, pp. 13–34.

Bohn, Ocke-Schwen, Catherine T. Best, Cinzia Avesani, and Mario Vayra. 2011. Perceiving through the lens of native phonetics: Italian and Danish listeners̀ perception of English consonant contrasts. In *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, China, August 17–21*. Edited by Wai Sum Lee and Eric Zee. Hong Kong: City University of Hong Kong, pp. 336–39.

Bradlow, Ann R., Reiko Akahane-Yamada, David B. Pisoni, and Yoh'ichi Tohkura. 1999. Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics* 61: 977–85.

Bradlow, Ann R., David B. Pisoni, Reiko Akahane-Yamada, and Yoh'ichi Tohkura. 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: 2299–310. [CrossRef]

Bundgaard-Nielsen, Rikke L., Catherine T. Best, Christian Kroos, and Michael D. Tyler. 2012. Second language learners' vocabulary expansion is associated with improved second language vowel intelligibility. *Applied Psycholinguistics* 33: 643–64. [CrossRef]

Bundgaard-Nielsen, Rikke L., Catherine T. Best, and Michael D. Tyler. 2011a. Vocabulary size is associated with second-language vowel perception performance in adult learners. *Studies in Second Language Acquisition* 33: 433–61. [CrossRef]

Bundgaard-Nielsen, Rikke L., Catherine T. Best, and Michael D. Tyler. 2011b. Vocabulary size matters: The assimilation of second-language Australian English vowels to first-language Japanese vowel categories. *Applied Psycholinguistics* 32: 51–67. [CrossRef]

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale: Lawrence Erlbaum Associates.

Cutler, Anne, Andrea Weber, and Takashi Otake. 2006. Asymmetric mapping from phonetic to lexical representations in second-language listening. *Journal of Phonetics* 34: 269–84. [CrossRef]

Darcy, Isabelle, Danielle Daidone, and Chisato Kojima. 2013. Asymmetric lexical access and fuzzy lexical representations in second language learners. *The Mental Lexicon* 8: 372–420. [CrossRef]

Faris, Mona M., Catherine T. Best, and Michael D. Tyler. 2016. An examination of the different ways that non-native phones may be perceptually assimilated as uncategorized. *Journal of the Acoustical Society of America* 139: EL1–EL5. [CrossRef] [PubMed]

Faris, Mona M., Catherine T. Best, and Michael D. Tyler. 2018. Discrimination of uncategorised non-native vowel contrasts is modulated by perceived overlap with native phonological categories. *Journal of Phonetics* 70: 1–19. [CrossRef]

Flege, James E. 1995. Second language speech learning: Theory, findings, and problems. In *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. Edited by Winifred Strange. Baltimore: York Press, pp. 233–76.

Flege, James E. 2003. Assessing constraints on second-language segmental production and perception. In *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*. Edited by Niels O. Schiller and Antje S. Meyer. Berlin: Mouton de Gruyter, pp. 319–55.

Flege, James E. 2009. Give input a chance! In *Input Matters in SLA*. Edited by Thorsten Piske and Martha Young-Scholten. Bristol: Multilingual Matters, pp. 175–90.

Flege, James E. 2019. A non-critical period for second-language learning. In *A Sound Approach to Language Matters—In Honor of Ocke-Schwen Bohn*. Edited by Anne M. Nyvad, Michaela Hejná, Anders Højen, Anna B. Jespersen and Mette H. Sørensen. Aarhus: Department of English, School of Communication & Culture, Aarhus University, pp. 501–41.

Flege, James E., Murray J. Munro, and Ian R. A. MacKay. 1995. Factors affecting strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America* 97: 3125–34. [CrossRef]

Flege, James Emil, Grace H. Yeni-Komshian, and Serena Liu. 1999. Age constraints on second-language acquisition. *Journal of Memory & Language* 41: 78–104.

Gibson, James J. 1966. *The Senses Considered as Perceptual Systems*. Boston: Houghton-Mifflin.

Gibson, James J. 1979. *The Ecological Approach to Visual Perception*. Boston: Houghton-Mifflin.

Goldstein, Louis, and Carol A. Fowler. 2003. Articulatory phonology: A phonology for public language use. In *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*. Edited by Niels O. Schiller and Antje S. Meyer. Berlin: Mouton de Gruyter, pp. 159–207.

Goto, Hiromu. 1971. Auditory perception by normal Japanese adults of the sounds "l" and "r". *Neuropsychologia* 9: 317–23. [CrossRef]

Guion, Susan G., James E. Flege, Reiko Akahane-Yamada, and Jesica C. Pruitt. 2000. An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *Journal of the Acoustical Society of America* 107: 2711–724. [CrossRef] [PubMed]

Hallé, Pierre A., Catherine T. Best, and Andrea Levitt. 1999. Phonetic vs. phonological influences on French listeners' perception of American English approximants. *Journal of Phonetics* 27: 281–306. [CrossRef]

Hattori, Kota, and Paul Iverson. 2009. English /r/-/l/ category assimilation by Japanese adults: Individual differences and the link to identification accuracy. *Journal of the Acoustical Society of America* 125: 469–79. [CrossRef] [PubMed]

Iverson, Paul, Patricia K. Kuhl, Reiko Akahane-Yamada, Eugen Diesch, Yoh'ichi Tohkura, Andreas Kettermann, and Claudia Siebert. 2003. A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition* 87: B47–B57. [CrossRef]

Jenkins, James J., Winifred Strange, and Linda Polka. 1995. Not everyone can tell a "rock" from a "lock": Assessing individual differences in speech perception. In *Assessing Individual Differences in Human Behavior: New Concepts, Methods, and Findings*. Edited by David J. Lubinski and René V. Dawis. Palo Alto: Davies-Black Publishing, pp. 297–325.

Kato, Misaki, and Melissa M. Baese-Berk. 2020. The effect of input prompts on the relationship between perception and production of non-native sounds. *Journal of Phonetics* 79: 100964. [CrossRef]

Lively, Scott E., John S. Logan, and David B. Pisoni. 1993. Training Japanese listeners to identify English /r/ and /l/: II. The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America* 94: 1242–55. [CrossRef]

Lively, Scott E., David B. Pisoni, and John S. Logan. 1992. Some effects of training Japanese listeners to identify English /r/ and /l/. In *Speech Perception, Production and Linguistic Structure*. Edited by Yoh'ichi Tohkura, Eric Vatikiotis-Bateson and Yoshinori Sagisaka. Washington: IOS Press, pp. 175–96.

Lively, Scott E., David B. Pisoni, Reiko A. Yamada, Yoh'ichi Tohkura, and Tsuneo Yamada. 1994. Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. *Journal of the Acoustical Society of America* 96: 2076–87. [CrossRef]

Logan, John S., Scott E. Lively, and David B. Pisoni. 1991. Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America* 89: 874–86. [CrossRef]

MacKain, Kristine S., Catherine T. Best, and Winifred Strange. 1981. Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied Psycholinguistics* 2: 369–90. [CrossRef]

McQueen, James M., Michael D. Tyler, and Anne Cutler. 2012. Lexical retuning of children's speech perception: Evidence for knowledge about words' component sounds. *Language Learning and Development* 8: 317–39. [CrossRef]

Miyawaki, Kuniko, James J. Jenkins, Winifred Strange, Alvin M. Liberman, Robert Verbrugge, and Osamu Fujimura. 1975. An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics* 18: 331–40.

Mochizuki, Michiko. 1981. The identification of /r/ and /l/ in natural and synthesized speech. *Journal of Phonetics* 9: 283–303. [CrossRef]

Norris, Dennis, James M. McQueen, and Anne Cutler. 2003. Perceptual learning in speech. *Cognitive Psychology* 47: 204–38. [CrossRef]

Olejnik, Stephen, and James Algina. 2003. Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods* 8: 434–47. [CrossRef]

Polka, Linda. 1995. Linguistic influences in adult perception of non-native vowel contrasts. *Journal of the Acoustical Society of America* 97: 1286–96. [CrossRef]

Reinisch, Eva, Andrea Weber, and Holger Mitterer. 2013. Listeners retune phoneme categories across languages. *Journal of Experimental Psychology: Human Perception and Performance* 39: 75–86. [CrossRef]

Shinohara, Yasuaki, and Paul Iverson. 2018. High variability identification and discrimination training for Japanese speakers learning English /r/–/l/. *Journal of Phonetics* 66: 242–51. [CrossRef]

Tyler, Michael D. 2019. PAM-L2 and phonological category acquisition in the foreign language classroom. In *A Sound Approach to Language Matters—In Honor of Ocke-Schwen Bohn*. Edited by Anne M. Nyvad, Michaela Hejná, Anders Højen, Anna B. Jespersen and Mette H. Sørensen. Aarhus: Dept. of English, School of Communication & Culture, Aarhus University, pp. 607–30.

Tyler, Michael D. 2021. Phonetic and phonological influences on the discrimination of non-native phones. In *Second Language Speech Learning: Theoretical and Empirical Progress*. Edited by Ratree Wayland. Cambridge: Cambridge University Press, pp. 157–74.

Tyler, Michael D., Catherine T. Best, Alice Faber, and Andrea G. Levitt. 2014a. Perceptual assimilation and discrimination of non-native vowel contrasts. *Phonetica* 71: 4–21. [CrossRef]

Tyler, Michael D., Catherine T. Best, Louis M. Goldstein, and Mark Antoniou. 2014b. Investigating the role of articulatory organs and perceptual assimilation in infants' discrimination of native and non-native fricative place contrasts. *Developmental Psychobiology* 56: 210–27. [CrossRef]

Weber, Andrea, and Anne Cutler. 2004. Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language* 50: 1–25. [CrossRef]

Werker, Janet F., and John S. Logan. 1985. Cross-language evidence for three factors in speech perception. *Perception & Psychophysics* 37: 35–44.

Winer, Benjamin J., Donald R. Brown, and Kenneth M. Michels. 1991. *Statistical Principles in Experimental Design*, 3rd ed. New York: McGraw-Hill.