Perception of nonnative tonal contrasts by Mandarin-English and English-Mandarin sequential bilinguals

I Lei Chan and Charles B. Chang[a)]
Department of Linguistics, Boston University, Boston, Massachusetts 02215, USA

(Received 14 September 2018; revised 12 July 2019; accepted 16 July 2019; published online 7 August 2019)

This study examined the role of acquisition order and crosslinguistic similarity in influencing transfer at the initial stage of perceptually acquiring a tonal third language (L3). Perception of tones in Yoruba and Thai was tested in adult sequential bilinguals representing three different first (L1) and second language (L2) backgrounds: L1 Mandarin-L2 English (MEBs), L1 English-L2 Mandarin (EMBs), and L1 English-L2 intonational/non-tonal (EIBs). MEBs outperformed EMBs and EIBs in discriminating L3 tonal contrasts in both languages, while EMBs showed a small advantage over EIBs on Yoruba. All groups showed better overall discrimination in Thai than Yoruba, but group differences were more robust in Yoruba. MEBs' and EMBs' poor discrimination of certain L3 contrasts was further reflected in the L3 tones being perceived as similar to the same Mandarin tone; however, EIBs, with no knowledge of Mandarin, showed many of the same similarity judgments. These findings thus suggest that L1 tonal experience has a particularly facilitative effect in L3 tone perception, but there is also a facilitative effect of L2 tonal experience. Further, crosslinguistic perceptual similarity between L1/L2 and L3 tones, as well as acoustic similarity between different L3 tones, play a significant role at this early stage of L3 tone acquisition.
 https://doi.org/10.1121/1.5120522

[TCB] Pages: 956–972

I. INTRODUCTION

Third language (L3) acquisition has received increasing attention in linguistic research and has been investigated in learner populations exemplifying various language backgrounds (e.g., Leung, 2003; Rothman and Cabrelli Amaro, 2010; Slabakova and del Pilar García Mayo, 2015). A central question in studies of L3 acquisition has been whether crosslinguistic influence or transfer from a previously-learned language comes from the bilingual learner's first language (L1), second language (L2), or both. Many studies have shown that the L1 or the L2 may be chosen as the primary source of transfer in L3 acquisition, on the basis of several different factors (see Hammarberg, 2010, for further discussion). These factors include (but are not limited to) prior language experience (Bardel and Falk, 2007), order of acquisition (Jin, 2009), proficiency and recency of language use (Cabrelli Amaro, 2013), typological similarity with the target L3 (Rothman and Cabrelli Amaro, 2010), and psychotypology (i.e., perceived similarity between languages; Foote, 2009; Leung, 1998).

However, L3 acquisition research has focused largely on Indo-European languages and few studies have investigated other widely-spoken types of languages, such as tone languages, in the context of multilingual development. Consequently, the degree to which the above-mentioned factors related to transfer in L3 acquisition may be generalized more broadly remains unclear. Given the need to understand patterns of transfer in L3 acquisition of lexical tone, a feature of over half of the world's languages (Yip, 2002), the study described in this paper investigated nonnative perception of L3 tonal contrasts by three types of sequential bilinguals: mirror-image groups of Mandarin-English bilinguals (i.e., L1 Mandarin-L2 English, L1 English-L2 Mandarin) as well as bilinguals with no tone language experience. In particular, we were interested in how previous knowledge of Mandarin tones would affect L3 acquisition of tones in a language that shares certain similarities with Mandarin (Thai) as compared to a language that shows more significant disparities (Yoruba).

In the following sections (Secs. II A–II C), we review previous work that developed theoretical models for L3 acquisition, discuss the literature on nonnative tone perception as well as the characteristics of the L3 tone inventories in this study, and then outline specific research questions regarding the effects of tone language (namely, Mandarin) transfer on the perception of L3 tones that align with the Mandarin tone system to different degrees.

II. BACKGROUND

A. Models of L3 acquisition

Formal linguistic inquiry into L3 acquisition has been shaped by three main theoretical models, which have addressed transfer at primarily a morphosyntactic level. Each model argues for a different criterion triggering the selection of a source language for transfer: utmost facilitation, order of acquisition, or typological similarity to the L3. The Cumulative Enhancement Model (Berkes and Flynn, 2012; Flynn *et al.*, 2004) suggests that previously acquired linguistic knowledge (from the L1 or L2) will transfer only when it is facilitative for L3 and subsequent (Ln) language learning. On the other hand, the L2 Status Factor Model

[a)]Electronic mail: cc@bu.edu

(Bardel and Falk, 2007; Falk and Bardel, 2011) posits that, on the basis of being the last-learned language (which is generally acquired in a more similar manner to the L3 than is the L1), the L2 serves exclusively as the source of transfer at the initial stage of L3 acquisition, regardless of whether such transfer is facilitative or non-facilitative. Finally, the Typological Primacy Model (TPM) (Rothman, 2011, 2015) argues that transfer is motivated by typological (i.e., structural) similarity between the L1/L2 and L3; however, because judgments of similarity may be based on a linguistic level that does not generalize to all levels, such transfer may or may not be facilitative. According to the TPM, after adequate L3 input at an early stage of L3 acquisition, either the L1 or L2 system is selected to transfer, based on four hierarchical levels of similarity with the L3: (1) lexicon, (2) phonology, (3) functional morphology, and finally (4) syntactic structure.

Although the Cumulative Enhancement Model and TPM both allow transfer from the L1 (based on different criteria), none of the above-mentioned models predict transfer to come consistently from the learner's L1. Several studies, however, have reported evidence of predominantly L1 influence in L3 acquisition (e.g., Hermas, 2014; Jin, 2009; Lozano, 2003; Na Ranong and Leung, 2009). For example, Hermas (2014) found that L1 Arabic-L2 French bilinguals transferred L1 patterns to L3 English, leading to both facilitative and non-facilitative instances of transfer, while Jin (2009) found, in a study of L1 Chinese-L2 English learners of L3 Norwegian, evidence of (non-facilitative) transfer from the L1 in spite of the fact that the L2 grammar in this case is more similar to that of the L3. These results suggest the possibility of an "L1 Status Factor" (see, e.g., Maimone, 2017; Neuser, 2017), which may arise due to the relative strength of the L1 in sequential bilinguals.

Whereas a possibly privileged status for the L1 has only recently been recognized as a factor in L3 morphosyntactic development, the L1 has long played a central role in explaining patterns in nonnative speech perception (e.g., Best and Tyler, 2007; Guion *et al.*, 2000). In the context of the current study, the framework of Automatic Selective Perception (ASP) (Strange, 2011) is particularly relevant. ASP proposes that L1 acquisition involves the development of L1-specific selective perception routines, which guide listeners toward automatized, efficient, and robust processing of the acoustic information relevant for identifying contrastive sound categories (i.e., phonemes) of the L1. These L1 selective perception routines are crucial to becoming a skilled listener of the L1 but may be unhelpful for processing a different language (which is likely to vary in terms of which acoustic cues are relevant for identifying phonemes). As a result, L2 learners are observed, over time, to develop new selective perception routines for processing the L2, which may also be drawn upon for processing the L1 (e.g., Carlson *et al.*, 2016). In the case of L3 perception, bilingual listeners may therefore bring L1 and/or L2 selective perception routines to the task. This additional complexity again raises the question of which language transfers to L3 perception. Although there is little research directly addressing this question, findings on L3 perception of Korean, Japanese, and Cantonese suggest that the L1, the L2, or both the L1 and L2 may influence L3 perception (Chang, 2018; Onishi, 2016; Qin and Jongman, 2016).

As discussed above, the main models of L3 acquisition (Cumulative Enhancement Model, L2 Status Factor Model, TPM) have been applied mostly to morphosyntax, yet their core logic can also be applied to speech sound acquisition. Indeed, several studies have focused on phonetics and phonology, reporting a diversity of results that have implications for these models (for reviews, see Cabrelli Amaro, 2012; Cabrelli Amaro and Wrembel, 2016; Wrembel, 2015). For example, some findings on L3 production (specifically, of voice onset time and accent) support the L2 Status Factor Model (Llama *et al.*, 2010; Wrembel, 2010), whereas other findings suggest more of a role for the L1, simultaneous transfer from both the L1 and the L2, or successive transfer from the L2 and then the L1 as L3 proficiency increases (Hammarberg and Hammarberg, 2009; Wrembel, 2012, 2015). Furthermore, a new line of research is pointing out that the L1 and L2 are unlikely to remain fixed during L3 learning (Cabrelli Amaro, 2017; Cabrelli Amaro and Rothman, 2010).

Thus, the current state of the science presents a complex picture with regard to predicting transfer in L3 phonology. With few exceptions, there is also a bias in the literature toward examining triads of related Indo-European languages, as well as a paucity of research on L3 perception. Consequently, we endeavored to explore models of L3 acquisition in the domain of speech perception by examining patterns of transfer in bilinguals who speak a non-Indo-European tone language (namely, Mandarin Chinese) when they begin to learn novel tones in an L3.

B. Nonnative tone perception and target systems

Prior research has shown that nonnative listeners whose L1 is non-tonal tend to have difficulty perceiving the contrastive tones of a tone language. For example, L1 English speakers were found not to perceive Mandarin or Thai tonal contrasts very accurately (Bent *et al.*, 2006; Burnham *et al.*, 1996). Nonnative listeners whose L1 is tonal, by contrast, may benefit from their tonal experience when perceiving novel tones, leading to a perceptual advantage over listeners from a non-tonal L1 background. Indeed, L1 Cantonese and Mandarin speakers were observed to be better than L1 English speakers at differentiating tones in their own language as well as those of Thai (Burnham *et al.*, 1996; Lee *et al.*, 1996; Wayland and Guion, 2004), an advantage that persisted under conditions of high stimulus variability (Chang *et al.*, 2017). Additionally, L1 Thai speakers were found to perceive artificial tone continua better than L1 English speakers (Burnham and Jones, 2002). However, L1 speakers of Hmong (a language with seven tones) perceived Mandarin tones less accurately than L1 English and Japanese speakers did (Wang, 2006, 2013), while L1 Burmese speakers and L1 (heritage) Cantonese speakers also, in certain cases, perceived Mandarin tones less accurately than listeners with no prior tonal experience (Tsukada and Kondo, 2019; Tsukada *et al.*, 2015). Although the explanation for this variation in the observed effect of prior tonal experience is unclear,

the conflicting results suggest that prior tonal experience can, but may not necessarily, be advantageous for the acquisition of nonnative tones. However, this set of findings, generally found with L2 as opposed to L3 learners, may not be generalizable to the case of bilingual listeners who have established more than one linguistic system, leaving open the question of whether bilinguals will show a facilitative or non-facilitative effect of prior tonal experience when they are at the beginning stages of learning a tonal L3.

Following from the TPM, it is possible that the nature of tone language transfer may depend on typological similarity with the target L3, which motivates the simultaneous investigation of multiple L3s that resemble the already-known tone language to different degrees. According to the TPM, a linguistic parser would first consider the lexical level in selecting a language to transfer, but in the case of unrelated tone languages with virtually no lexical overlap, the parser would move on to the phonological level, where there is great typological diversity in the nature of tone systems. Crosslinguistic variation occurs, for example, in the specific inventory of tones, in the use of particular phonetic features (e.g., pitch height, pitch contour) for classifying tones, and in tones' distinctive functions (e.g., lexical contrast, grammatical contrast). The four main languages/language types included in the present study thus overlap to different degrees in their use of pitch variation.

As for the non-target languages (already known to learners), the first type was Mandarin Chinese, a contour tone language with four main tones, one static tone and three dynamic tones. Traditionally, these tones are labeled Tone 1 (hereafter, M1[1]) "high level" [a^{55}], Tone 2 (M2) "mid rising" [a^{35}], Tone 3 (M3) "low dipping" [a^{214}], and Tone 4 (M4) "high falling" [a^{51}] (Chao, 1968), and transcribed using a five-point scale (1 = low end of a talker's pitch range, 5 = high end; see Chao, 1930) to indicate the pitch contour. The Mandarin tones are shown in Fig. 1, which plots their f_0 contours in terms of speaker-normalized f_0 (T) from acoustic analyses of all of the Mandarin monosyllabic test stimuli (see Sec. III B; analysis methodology from Chang and Yao, 2016). Mandarin is classified as a contour tone language because, as evident in Fig. 1, it uses primarily pitch contour

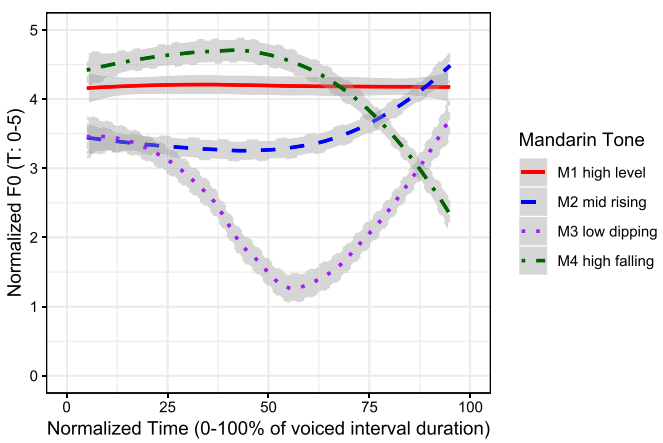


FIG. 1. (Color online) Mandarin tone contours (in terms of T), as produced in the test stimuli. The curve for each tone represents the local smoothing function fit to all of the acoustic data for that tone (i.e., `method` = "`loess`" in R); each shaded area represents the 95% confidence interval around this curve.

(i.e., the pattern of pitch change over time) to differentiate the tones, which create lexical contrasts (e.g., /ta^{55}/ "carry," /ta^{35}/ "reach," /ta^{214}/ "hit," /ta^{51}/ "large"; So and Best, 2010); that is, Mandarin tones can each be distinguished by their pitch contour, and there are no two tones that contrast primarily in terms of pitch level. The second type of non-target language (again, already known to learners), which effectively served as a control language within each bilingual profile, comprised non-tonal (in particular, intonation) languages, such as English, other Germanic languages, and Romance languages (e.g., Spanish, French); these languages differ crucially from tone languages in not containing lexically contrastive tones.[1]

As for the target L3s, we selected two tone languages, Thai and Yoruba, based on the degree of typological similarity between their tone system and that of Mandarin. Thai has been described as a contour tone language like Mandarin, with a five-tone inventory including three static tones and two dynamic tones (Abramson, 1962; Gandour and Harshman, 1978). Following the conventional tone ordering of native speakers (which differs from the ordering often seen in the linguistic literature; cf. Gandour and Harshman, 1978; James, 1923; Yang, 2019), these tones are referred to here as Tone 1 (hereafter, H1) "mid" [a], Tone 2 (H2) "low" [à], Tone 3 (H3) "falling" [â], Tone 4 (H4) "high" [á], and Tone 5 (H5) "rising" [ă]. By contrast, Yoruba has been described as a register tone language, with a three-tone inventory consisting of register, or level, tones (Courtenay, 1968; Gandour and Harshman, 1978). Again, following the conventional ordering of native speakers, these are referred to here as Tone 1 (Y1) "low" [à], Tone 2 (Y2) "mid" [a], and Tone 3 (Y3) "high" [á]. As in Mandarin, different tones in both Yoruba and Thai create lexical contrasts (e.g., Yoruba: /lɔ̀/ 'to grind', /lɔ/ 'to go', /lɔ́/ 'lukewarm', Thai: /k^{h}a:/ 'to be stuck', /k^{h}à/ 'a kind of spice', /k^{h}â:/ 'to kill', /k^{h}á:/ 'to engage in trade', /k^{h}ă:/ 'leg'; Gandour and Harshman, 1978). However, unlike Mandarin, only Yoruba is known to use tone grammatically (Agwuele, 2005; Bamgbose, 1967; Lamidi, 2003). Thus, with an inventory of register tones, which is used linguistically in a different way than in Mandarin, Yoruba's tone system can be said—at least in principle—to diverge to a greater degree from Mandarin's tone system than does Thai's.

Given the diachronic instability of tone (Remijsen, 2016) as well as recent findings of sound change in progress in Thai tones (Teeranon, 2008; Thepboriruk, 2009), we examined the specific manifestations of the target L3 tone systems in this study by carrying out acoustic analyses of all of the monosyllabic test stimuli for both L3s (see Sec. III B) in the same manner as for Mandarin, and the results are shown in Fig. 2 (f_0 contours) and Table I (durations). As can be seen in Fig. 2, all of the Yoruba and Thai tones show some degree of pitch movement; that is, although some of these tones have been previously described as level tones, none have as level a shape as M1 in Mandarin, although H1 comes close. Thus, to classify the L3 tone systems in a more nuanced manner, every possible tone contrast within each language was classified as one of two types: a contrast in pitch contour or a contrast in (primarily) pitch register (i.e., level). Considering the two languages in this more fine-grained,

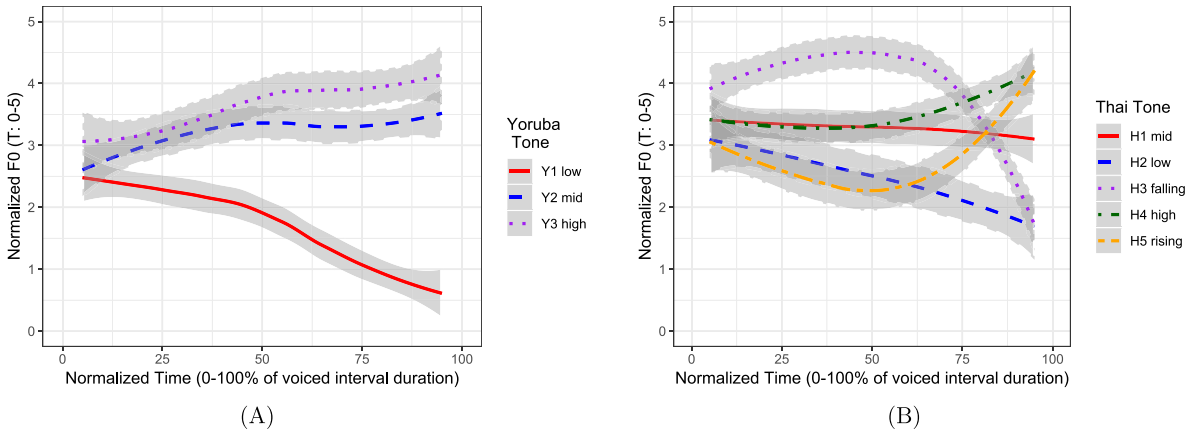


FIG. 2. (Color online) Yoruba (A) and Thai (B) tone contours (in terms of *T*), as produced in the test stimuli. The curve for each tone represents the local smoothing function fit to all of the acoustic data for that tone (i.e., `method` = "`loess`" in R); each shaded area represents the 95% confidence interval around this curve.

contrast-by-contrast manner revealed that the Yoruba system consists of two contour contrasts (Y1-Y2, Y1-Y3) and one register contrast (Y2-Y3), while the Thai system consists solely of contour contrasts; that is, none of Thai's ten possible tone contrasts is a clear register contrast.[2] By comparison, Mandarin's tone system also consists of only contour contrasts, with none of its six tone contrasts qualifying as a register contrast. Furthermore, whereas Mandarin shows large durational differences among some tones (e.g., M4 is much shorter than M3; see Table I), Yoruba and Thai show relatively small durational differences among their tones. Therefore, it is reasonable to assume that pitch cues are indeed crucial to distinguishing tones in both L3s, and that, in a continuum of tone systems ranging from the maximally "contour-like" system consisting of only contour contrasts to the maximally "register-like" system consisting of only register contrasts, Thai's system is closer to the prototypical contour system of Mandarin than is Yoruba's system.

C. Research questions and predictions

The current study explored three questions pertaining to L3 perception of Thai and Yoruba tones. First, is previous tonal experience facilitative in L3 tone perception (**Q1**)? Second, whatever the effect of prior tonal experience, is the effect observed (or, alternatively, larger) when the tonal experience is from the L1 or from the L2 (**Q2**)? Third, does typological similarity (i.e., similarity in terms of one or more linguistic structural features) and/or perceptual similarity (i.e., similarity in terms of how languages are perceived by the listener, either with respect to a specific structure or at a holistic level) between bilinguals' tonal L1/L2 (i.e., Mandarin) and the target tonal L3 affect the perception of L3 tonal contrasts (**Q3**)? The target tonal L3s selected for this study were Yoruba and Thai, for two reasons: (1) the absence of lexical overlap between Mandarin and either language, which allows for an examination of the role of typological similarity at the phonological level (cf. the TPM),[3] and (2) the clear disparity between the two languages in terms of their typological similarity (with respect to tone system) to Mandarin, as described in Sec. II B.

One objective of this study was to identify which of the theoretical approaches to L3 acquisition can best explain L3 acquisition of tone by sequential bilinguals who know one, and only one, tone language. To this end, several possible outcomes in L3 tone perception were considered in light of whether they would be predicted by these theories (see Table II). One possible outcome is Mandarin-speaking bilinguals, both L1 Mandarin-L2 English (MEBs) and L1 English-L2 Mandarin (EMBs), outperforming L1 English speakers who are bilingual in an intonation/non-tonal language (EIBs), due to a facilitative effect of prior tonal experience. This outcome would be consistent with predictions of the Cumulative Enhancement Model, but also of the TPM (assuming that, in the absence of lexical overlap, tone

TABLE I. Durations (in ms) of contrastive lexical tones in Mandarin, Yoruba, and Thai. Means and standard deviations are over the nine monosyllabic test stimuli (3 speakers × 3 vowel contexts) used in the perception experiments.

Mandarin			Yoruba			Thai		
tone	*M*	*SD*	tone	*M*	*SD*	tone	*M*	*SD*
M1	447	65	Y1	367	70	H1	551	85
M2	469	73	Y2	405	85	H2	545	95
M3	572	118	Y3	418	86	H3	544	62
M4	322	35				H4	551	95
						H5	557	97

TABLE II. Possible outcomes (i.e., relative performance in L3 perception) predicted by models of later language learning ("Y" = outcome predicted). The four models are the Cumulative Enhancement Model (CEM), L2 Status Factor (L2SF), TPM, and ASP; the three groups shown in each outcome are L1 Mandarin-L2 English (MEB), L1 English-L2 Mandarin (EMB), and L1 English-L2 intonational (EIB).

Outcome (> "better than"; = "as good as")	CEM	L2SF	TPM	ASP
MEB = EMB > EIB	Y	—	Y	—
EIB > MEB = EMB	—	—	Y	—
EMB > MEB = EIB	—	Y	—	—
MEB = EIB > EMB	—	Y	—	—
MEB > EMB > EIB	—	—	—	Y

languages are deemed more similar to each other than are tone and intonation languages). The reverse result (i.e., EIBs outperforming MEBs and EMBs), which could arise if prior tonal experience is non-facilitative, would also be consistent with predictions of the TPM, but not of the Cumulative Enhancement Model. Another possible result is EMBs patterning distinctly from MEBs and EIBs, either outperforming them due to a facilitative effect of prior tonal experience in the L2 specifically or underperforming them due to a non-facilitative effect. These outcomes would be consistent with predictions of the L2 Status Factor, but of no other theory. Last, all three groups could pattern differently from each other. For example, under the dominant influence of the L1, MEBs could outperform EMBs and EIBs due to greater attunement to pitch in their L1 selective perception routines facilitating L3 tone perception, with EIBs performing the worst because none of their selective perception routines (L1 or L2) involve attunement to pitch on the timescale of tones. This outcome would follow from ASP, but would be inconsistent with predictions of the other theories.

Given previous evidence of prior tonal experience being advantageous for acquiring nonnative tones (e.g., Qin and Jongman, 2016; Wayland and Guion, 2004) as well as the theory that L1 experience leads to a "neural commitment" to L1-specific auditory patterns (Kuhl, 2000; Zhang *et al.*, 2005), we had a basis for formulating three specific predictions about how the three groups would fare in L3 tone perception. Our first prediction (**P1**) was that prior tonal experience would be overall advantageous for L3 tone perception, while our second prediction (**P2**) was that the advantage of tonal experience would be greater when coming from the L1 than the L2. Thus, we expected that, of the three groups, the L1 Mandarin group (MEBs) would show the greatest sensitivity to L3 tonal contrasts across different L3s. Our third prediction (**P3**) was that, based on the greater typological similarity in tone system (or, possibly, greater holistic perceptual similarity) between Thai and Mandarin compared to Yoruba and Mandarin, prior tonal experience would transfer to a greater degree to Thai, resulting in Mandarin-speaking bilinguals' advantage over non-tonal bilinguals being greater in Thai. In short, we predicted that prior tonal experience would be facilitative, that this facilitation would be stronger in the case of L1 transfer, and that typological similarity would play a crucial role in modulating this transfer.

Apart from typological similarity at the level of a specific feature, structure, or pattern, crosslinguistic similarity can also be conceptualized in other ways, including perceptual similarity. In fact, perceptual similarity between phonetic elements of different languages figures prominently in the literature on nonnative speech perception and is the basis of a widely tested theory, the Perceptual Assimilation Model (Best, 1994). In brief, the Perceptual Assimilation Model posits that nonnative listeners are inclined to assimilate nonnative sounds to similar native sounds, and that different patterns of crosslinguistic assimilation lead to different patterns of discrimination. According to this model, if nonnative sounds are assimilated to distinct native categories ("Two Category" assimilation), they will be discriminated well, whereas if the nonnative sounds are assimilated to the same category ("Single Category" assimilation), they will be discriminated poorly (see, e.g., Best and Strange, 1992; Guion *et al.*, 2000). On the other hand, if the nonnative sounds are assimilated to a single category but with a different goodness-of-fit ("Category Goodness" assimilation), discrimination will be intermediately difficult. Given the potential explanatory power of perceptual similarity, we collected data not only on discrimination, but also on similarity, both between tones and between languages, in order to test the role of perceptual similarity in L3 tone perception. Our fourth prediction (**P4**) was thus that L3 tonal contrasts would be variably difficult to discriminate, with difficulty for Mandarin speakers being correlated with perceived similarity to the same previously-acquired (i.e., Mandarin) tone.

Our final prediction was based in (psycho)acoustic similarity, due to its role in accounting for perceptual variation across tone contrasts even among native listeners of the target language. In short, different tone contrasts are variably difficult to discriminate for L1 listeners (including those who know no other tone language and thus should not be influenced by any other tone systems) as well as for computers, which can typically be attributed to variation in the objective (i.e., acoustic) similarity of different tone pairs. For example, Y2 (mid) and Y3 (high) are more confusable for an artificial neural network trained on Yoruba tones than is either of the other tone pairs (Ọdẹlóbí, 2008), consistent with Y2 and Y3's close acoustic proximity [see Fig. 2(B) and Table I].[4] As for Thai, H1 (mid) and H2 (low), which overlap in their initial f_0 level and show the same directionality of f_0 movement, have been shown to be more confusable for native Thai listeners than other tone pairs (Abramson, 1976). Along the same lines, H2 (low) and H3 (falling), as well as H4 (high) and H5 (rising), are also acoustically similar tone pairs because they overlap in f_0 level at one or both ends of their contour and show the same directionality of f_0 movement for half or more of their duration. Thus, our fifth prediction (**P5**) was that L3 tone discrimination would be significantly influenced by acoustic similarity, such that even bilinguals with no tone language experience (EIBs) would find certain tone pairs (e.g., Y2-Y3, H1-H2) more difficult to discriminate than others.

III. METHODS

A. Participants

All listener participants were recruited from the Greater Boston metropolitan area. To be eligible for the study, participants had to identify as bilingual, with their two languages being Mandarin and English (in either order of acquisition) or the L1 being English and the L2 an intonation (i.e., non-tonal) language. Participants included in the final sample additionally reported no history of hearing, speech, or language impairments, no recent musical training, and no prior exposure to Yoruba or Thai. They gave informed consent at the beginning of the study and were paid for their participation.

There were thus three groups of bilingual participants: L1 Mandarin-L2 English bilinguals (MEBs), L1 English-L2 Mandarin bilinguals (EMBs), and L1 English-L2 intonational bilinguals (EIBs). With the exception of three switched-dominance EIBs, all were sequential bilinguals with an age of

acquisition (AOA) for L2 of four years or later. The MEB group consisted of 20 native speakers of Mandarin [eight males; $M_{age} = 24.5$ years, standard deviation (*SD*) 1.9] born and raised in China who had learned English as an L2 ($M_{AOA} = 7.5$ years, *SD* 2.4). Half of the MEBs also reported knowledge of an additional language besides English (e.g., French, Russian); however, with the exception of Cantonese for one participant, none of these languages were tonal.[5] The EMB group comprised 18 native speakers of English (seven males; $M_{age} = 20.7$ years, *SD* 2.3) raised primarily in the US who had learned Mandarin as an L2 ($M_{AOA} = 13.5$ years, *SD* 4.1). About half of the EMBs reported knowledge of an additional language besides Mandarin, but none of these languages were tonal. The EIB group consisted of 18 native speakers of English (four males; $M_{age} = 25.8$ years, *SD* 6.2) raised primarily in the US who had acquired an intonation language as an L2 ($M_{AOA} = 11.7$ years, *SD* 6.8). The L2s represented in the EIB group were Dutch, French, German, and Spanish. Most EIBs reported knowledge of a language beyond their specific L2 as well, but again none of these languages were tonal. Further information about the additional languages known by participants, including self-reported proficiency levels, is provided in the open-access datasets online.[6]

All groups had received extensive exposure to their L2 (on average, 6.5 years or more), rated their L2 proficiency as intermediate or higher, and reported using both of their languages frequently. Data on language proficiency and use were gathered via a detailed background questionnaire (adapted from an instrument measuring English-Mandarin bilinguals' language dominance and recency of language use; see Lim *et al.*, 2008, pp. 405–410), which was completed after both perception experiments. The full questionnaire is publicly accessible.[7,8,9]

To provide an objective measure of their L2 proficiency, participants were also given a lexically-based test of their L2 knowledge (LexTALE or an adaptation thereof). The L2 English proficiency of MEBs was assessed via LexTALE (Lemhöfer and Broersma, 2012), which tests English lexical knowledge by asking users to identify the wordhood status of 60 items (40 real words, 20 non-words). The L2 Mandarin proficiency of EMBs was evaluated via a Mandarin version of LexTALE (LEXTALE_CH; Chan and Chang, 2018). As for the EIB group, EIBs represented a variety of non-tonal L2s: Dutch, French, German, and Spanish. The L2 proficiency of EIBs was therefore assessed via the Spanish version of LexTALE (LEXTALE_ESP; Izura *et al.*, 2014), the French version (LEXTALE_FR; Brysbaert, 2013), or the Dutch and German versions developed by Lemhöfer and Broersma. The scores on these tests (converted to the same 100-point scale using the calculation in Izura *et al.*, 2014, p. 58) further suggested that all groups comprised proficient bilinguals. MEBs, EMBs, and EIBs obtained mean scores of 64.3 (*SD* 9.7), 65.6 (*SD* 7.9), and 63.5 (*SD* 13.5), respectively, on the test for their L2. These score levels, which all indicate an "upper intermediate" level of proficiency in the L2, were not significantly different [Welch-corrected two-sample $|t|s < 0.526$, $ps > 0.05$].

B. Stimuli

All stimuli for the perception experiments were audio-recorded by three different native speakers of the target language in order to be able to incorporate talker variability into the experiments. The Yoruba talkers were two females and a male ($M_{age} = 38$ years) born and raised in Nigeria (specifically, Lagos, Ikere-Ekiti, or Ondo); the Thai talkers, two females and a male ($M_{age} = 26$ years) born and raised in Thailand (specifically, Bangkok); and the Mandarin talkers (for the Mandarin stimuli used in the similarity rating experiment), two females and a male ($M_{age} = 25$ years) born and raised in mainland China (specifically, Liaoning, Jiangsu, or Hunan). All talkers spoke a variety of the language close to the standard dialect, and were proficient in English as well; thus, their speech might be considered representative of the input in these languages that learners based in the US, such as the listener participants in this study, would generally be exposed to the most.

Stimulus recording took place in a sound-attenuated booth in the US using an AKG C520 head-worn condenser mic and a MacBook laptop running Praat (Boersma and Weenink, 2016). The list of stimuli (consonant-vowel monosyllables, along with short passages; see Secs. III C 1 and III C 2) was presented to the talker in the booth via a PowerPoint slideshow, which primed target tones with real words in the respective language to make it clear which tone was intended in any given item. Target syllables were shown in International Phonetic Alphabet (IPA) symbols (including tone marks) along with additional clarification (e.g., "lá - high tone," "là - low tone"); target passages were shown in the respective orthography. Talkers were instructed to produce each stimulus three times, and one of these tokens was later selected (on the basis of the general clarity of the articulation, as judged by the first author in post-recording auditory inspection) and then amplitude-normalized (for peak amplitude) for use in the experiments.

In order to check for disparities in acoustic variability (in particular, pitch variability; see Table I for data on durational variability) among the three stimulus languages, three pitch parameters were analyzed in the monosyllabic items: (i) the low point (f_0 minimum) of a given tone contour, (ii) the high point (f_0 maximum) of the contour, and (iii) the distance in pitch traversed by the contour (f_0 range). The results of this analysis are summarized in Table III. In short, the amount of variability differed by tone and by f_0 parameter, but overall there appeared to be more variability (as measured by the

TABLE III. Variability in f_0 parameters (in terms of standard deviations from the mean, in Bark) across stimulus languages, by tone. Each standard deviation is over the nine monosyllabic test stimuli (3 speakers × 3 vowel contexts) for that tone used in the perception experiments.

Mandarin tone	f_0 min	f_0 max	f_0 range	Yoruba tone	f_0 min	f_0 max	f_0 range	Thai tone	f_0 min	f_0 max	f_0 range
M1	0.45	0.44	0.06	Y1	0.49	0.41	0.17	H1	0.72	0.70	0.06
M2	0.26	0.36	0.21	Y2	0.46	0.59	0.33	H2	0.55	0.66	0.36
M3	0.29	0.18	0.41	Y3	0.41	0.33	0.15	H3	0.51	0.59	0.44
M4	0.53	0.47	0.83					H4	0.66	0.78	0.16
								H5	0.57	0.56	0.23

average standard deviation over all tones) in the Thai stimuli than in the Yoruba stimuli; this was true with respect to f_0 minimum ($M_{SD} = 0.60$ Bark in Thai vs $M_{SD} = 0.45$ Bark in Yoruba; cf. $M_{SD} = 0.36$ Bark in Mandarin), f_0 maximum ($M_{SD} = 0.66$ Bark in Thai vs $M_{SD} = 0.44$ Bark in Yoruba; cf. $M_{SD} = 0.36$ Bark in Mandarin), and, to a lesser extent, f_0 range ($M_{SD} = 0.25$ Bark in Thai vs $M_{SD} = 0.22$ Bark in Yoruba; cf. $M_{SD} = 0.38$ Bark in Mandarin). As shown in Table I, there was also a tendency for durational variability to be greater in Thai ($M_{SD} = 87$ ms) than Yoruba ($M_{SD} = 80$ ms), as well as Mandarin ($M_{SD} = 73$ ms). Thus, given that acoustic variability, including both intra- and inter-talker variability, typically increases the difficulty of a perceptual task, this disparity could make the discrimination task easier for Yoruba than for Thai (however, as shown in Sec. IV A, this possibility is not borne out).

C. Procedure

Listeners completed two perception experiments in a quiet room: an oddity discrimination experiment (Experiment 1) and then a similarity rating experiment (Experiment 2). The experiments were run in OpenSesame 3.1.6 (Mathôt *et al.*, 2012) using either a Lenovo or iMac desktop computer, a Cedrus 7-button response pad (RB-740), and a pair of studio-quality binaural headphones. At the beginning of each experiment, both oral instructions (in-person explanation) and written (on-screen) instructions for the tasks were given to listeners in their native language. In particular, they were instructed to listen to the tonal contrasts carefully and to respond as quickly as possible. The experiments were completed in the following sequence, with optional intervening breaks: Experiment 1 (Yoruba section, Thai section), Experiment 2 (Yoruba section, Thai section). Since Experiment 2 involved Mandarin stimuli, it was completed after Experiment 1 so as to postpone targeted exposure to Mandarin tones until after the task not specifically involving Mandarin. Further, within each experiment, the shorter section was ordered before the longer section so as to decrease the chance of listener fatigue (by introducing a break opportunity earlier in the experiment); this meant that the Yoruba section of an experiment always occurred before the Thai section due to Yoruba's smaller tone inventory.

1. Experiment 1: Tone discrimination

The stimuli for Experiment 1 included three different consonant-vowel (CV) syllables, which were combined with the five Thai and three Yoruba tones to create a total of 24 target items. The target syllables combined the lateral /l/ with the vowels /e/, /a/, and /o/. These syllables were chosen for several reasons. First, CV is an unmarked syllable type across languages, allowed in both target L3s and the two L1s represented in the study sample (English, Mandarin). Second, the constituent segments are typologically common and found in all four languages with a broadly similar phonetic quality; the use of such common segments was meant to avoid introducing any processing burden associated with marked segments, thereby guiding listeners toward focusing on the nonnative tonal contrasts. Third, compared to other possible syllables (e.g., /li/), the target syllables /la le lo/ reduce the potential for semantic interference from similar words in English and Mandarin: General American English, where the tense mid vowels are diphthongal, has no lexical items that overlap perfectly since the target vowels are monophthongal, whereas Mandarin has only a few items that overlap segmentally (e.g., /la^{55}/ "emotional marker," /la^{214}/ "horn," /la^{51}/ "spicy").

In this experiment, listeners completed a speeded (i.e., respond as quickly as possible) four-alternative forced choice (4AFC) oddity discrimination task (Flege, 2003; Flege and MacKay, 2004), which used the full set of Yoruba and Thai tones. On each trial, three auditory stimuli produced by different talkers were presented in sequence, separated by an inter-stimulus interval (ISI) of 1.2 s. After the three items were played, listeners had to identify whether the items had the same tone or whether one had a different tone from the other two. They were instructed to press the button on the response box marked with the number ("1," "2," or "3") indicating the serial position of the odd item out (if relevant) and otherwise to press "4" to indicate that the items had the same tone.

There were two types of trials, which each presented a three-item auditory sequence: "change" and "no change" trials. The "change" trials contained an odd item that varied in tone from the other two items (which had the same tone), with the odd item occurring with equal frequency in all possible serial positions. The "change" trials comprised 86% and 92% of all trials, respectively, for Yoruba and Thai; the remainder comprised "no change" trials. The stimuli on each trial thus contained the same segments and usually, but not always, varied in tone.

The incorporation of talker variability within each trial and the relatively long ISIs were intended to encourage discrimination of the tones at a more abstract level (Flege, 2003) rather than purely at an acoustic level; this also made the task more difficult. In addition, the inclusion of both "change" and "no change" trials was based on the view (e.g., Guenther *et al.*, 1999; Kuhl, 1980) that the formation of a new phonetic category increases sensitivity to differences between members of the new category and other native categories, but decreases sensitivity to differences among members of the new category. The "change" trials, therefore, were meant to test listeners' ability to discern categorically different tones, while the "no change" trials tested their ability to ignore audible, but phonemically irrelevant, within-category variation.

Experiment 1 consisted of two sections. The Yoruba section contained an initial practice block of five trials (a subset of the test trials) without feedback, and then a randomized test block comprising a total of 63 trials (54 "change" trials = 3 tone contrasts × 2 possible oddball tones × 3 possible oddball positions × 3 syllables; 9 "no change" trials = 3 tones × 3 syllables). The Thai section contained an initial practice block of five trials (again, a subset of the test trials) without feedback and then a randomized test block comprising a total of 195 trials (180 "change" trials = 10 tone contrasts × 2 possible oddball tones × 3 possible oddball positions × 3 syllables; 15 "no change" trials = 5 tones × 3 syllables). All trial types were distributed evenly across

blocks and across all possible talker orders. With an ISI and inter-trial interval of 1.2 s each, Experiment 1 took approximately 25 min in all and produced two measures: percent accuracy (i.e., likelihood of a correct response) and mean reaction time (RT) for correct responses.

2. Experiment 2: Similarity rating

The stimuli for Experiment 2, which involved ratings of crosslinguistic similarity, included the Yoruba and Thai stimuli used in Experiment 1, as well as additional Mandarin stimuli constructed similarly to the Yoruba and Thai stimuli. To create the Mandarin stimuli, the onset /l/ and vowel nuclei /ei/, /a/, and /ou/ were combined with the four Mandarin tones for a total of 12 target items. For the latter part of Experiment 2, which examined global (holistic) perceptual similarity between Yoruba/Thai and Mandarin, the stimuli were longer audio passages from Aesop's fable "The North Wind and the Sun" translated into Mandarin, Yoruba, and Thai.

In this experiment, listeners completed a perceptual similarity rating task in which they heard tokens spoken by a single talker from each target L3 juxtaposed with a token of Mandarin (spoken by a talker of the same gender), in the sequence of L3-Mandarin. Each trial therefore presented two stimuli, and participants had to rate the tone in the first (L3) stimulus in terms of its similarity to the tone in the second stimulus on a 1–7 scale (1 = very different, 7 = very similar). The Yoruba section contained an initial practice block of three trials and then a randomized test block of 36 trials (3 Yoruba tones × 4 Mandarin tones × 3 syllables). The Thai section contained an initial practice block of three trials and then a randomized test block of 60 trials (5 Thai tones × 4 Mandarin tones × 3 syllables).

After all tone similarity ratings were given, the experiment then proceeded to a final part, consisting of two trials total, in which longer passages of continuous speech (as opposed to monosyllables) were evaluated. On these last two trials, a passage of the L3 (lasting about 50–70 s) was played, followed by a slightly shorter passage of Mandarin (lasting about 40–50 s, again spoken by a talker of the same gender as the L3 talker), and participants had to rate the overall perceptual similarity between the L3 and Mandarin on the same 1–7 scale. Experiment 2 took approximately 15 min in all and produced two types of measures: tone similarity rating (between paired tones) and holistic similarity rating (between languages, based on paired passages).

IV. RESULTS

A. Experiment 1: Tone discrimination

The likelihood of an accurate response in Experiment 1 was analyzed with logistic mixed-effects regression, using `glmer()` in the `lme4` package (Bates *et al.*, 2019) in R (R Development Core Team, 2018). All data from Experiment 1, as well as Experiment 2, are publicly accessible.[10] The initial omnibus model of accuracy contained random intercepts for Participant and Item (i.e., the specific sequence of auditory stimuli presented on a given trial) and four simple fixed effects: treatment-coded effects for Group (EIB, EMB, MEB; reference level = EIB), Language (Thai, Yoruba; reference level = Thai), and Trial Type (change, no change; reference level = change), and a continuous effect for Order (i.e., the ordinal position of a trial within the experimental session). In addition, this model included all eleven possible interactions among the fixed predictors.

An analysis of variance (ANOVA) on this initial model [using `Anova()` in the `car` package; Fox *et al.*, 2018] revealed that most of the interactions were not significant [$\chi^2 < 3.297$, $p > 0.05$], so to aid interpretation of the fixed-effect coefficients, all non-significant interactions involving the non-critical (i.e., control) predictor Order were pruned from the initial model, leaving four interactions in the final model. An ANOVA on this model showed significant main effects of Group [$\chi^2(2) = 13.081$, $p = 0.001$], Language [$\chi^2(1) = 18.763$, $p < 0.0001$], Trial Type [$\chi^2(1) = 9.064$, $p = 0.003$], and Order [$\chi^2(1) = 20.962$, $p < 0.0001$]; in addition, both the Group × Language [$\chi^2(2) = 8.469$, $p = 0.014$] and Group x Trial Type [$\chi^2(2) = 23.730$, $p < 0.0001$] interactions were significant. The Language × Trial Type interaction was not significant [$\chi^2(1) = 0.077$, $p = 0.781$], while the Group × Language × Trial Type interaction was marginal [$\chi^2(2) = 4.630$, $p = 0.099$]. Table IV summarizes the fixed-effect coefficients in this model. Note that Order had a small, but significant, positive effect, but there was a separate effect of Language and no significant interaction between Order and Language. In other words, the Language effect discussed below was statistically distinct from the effect of Yoruba trials coming before Thai trials; therefore, we focus on the Language effect below because this is the critical effect for our research questions.

The variation of discrimination accuracy across groups and languages is depicted in Fig. 3. As can be seen in Fig. 3, accuracies were lowest for EIBs ($M = 50\%$), higher for EMBs ($M = 56\%$), and highest for MEBs ($M = 65\%$); in addition, accuracies were, for all groups, higher on Thai ($M_{EIB} = 60\%$, $M_{MEB} = 71\%$, $M_{EMB} = 64\%$) than Yoruba ($M_{EIB} = 41\%$, $M_{MEB} = 60\%$, $M_{EMB} = 47\%$), and all groups showed higher accuracies on "no change" trials ($M_{EIB} = 66\%$, $M_{MEB} = 85\%$, $M_{EMB} = 64\%$) compared to "change" trials ($M_{EIB} = 54\%$, $M_{MEB} = 67\%$, $M_{EMB} = 60\%$).[11] These patterns were reflected in the main effects of Group, Language, and Trial Type from above. The coefficients of the main model (Table IV) showed that, on "change" trials in Thai, MEBs were significantly more likely to be accurate than EIBs [$\beta = 0.668$, $z = 3.105$, $p = 0.002$], whereas EMBs were not [$\beta = 0.329$, $z = 1.492$, $p = 0.136$]; further, EIBs were less likely to be accurate on Yoruba than Thai "change" trials [$\beta = -0.910$, $z = -4.247$, $p < 0.0001$], and more likely to be accurate on "no change" than "change" trials in Thai [$\beta = 1.139$, $z = 3.032$, $p = 0.002$]. Additionally, the interaction coefficients indicated that the decrement in accuracy observed on Yoruba "change" trials for EIBs was significantly smaller for MEBs [$\beta = 0.246$, $z = 2.044$, $p = 0.041$], that the increment in accuracy observed on Thai "no change" trials for EIBs was significantly smaller for EMBs [$\beta = -0.674$, $z = -3.001$, $p = 0.003$], and that the combined effect of switching to Yoruba and to "no change" trials was significantly more positive for MEBs compared to EIBs [$\beta = 0.797$, $z = 2.115$, $p = 0.034$].

TABLE IV. Fixed-effect terms in the main logistic mixed-effects model of accuracy in Experiment 1 (tone discrimination). Significance codes: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Predictor	β	SE	z	p
(Intercept)	0.200	0.186	1.078	0.281
Group: MEB	0.668	0.215	3.105	0.002**
Group: EMB	0.329	0.220	1.492	0.136
Language: Yoruba	−0.910	0.214	−4.247	<0.0001***
Trial Type: no change	1.139	0.376	3.032	0.002**
Order	0.002	0.0004	4.578	<0.0001***
Group: MEB × Language: Yoruba	0.246	0.120	2.044	0.041*
Group: EMB × Language: Yoruba	0.069	0.123	0.558	0.577
Group: MEB × Trial Type: no change	0.081	0.244	0.332	0.740
Group: EMB × Trial Type: no change	−0.674	0.224	−3.001	0.003**
Language: Yoruba × Trial Type: no change	−0.551	0.615	−0.896	0.370
Group: MEB × Language: Yoruba × Trial Type: no change	0.797	0.377	2.115	0.034*
Group: EMB × Language: Yoruba × Trial Type: no change	0.469	0.350	1.340	0.180

To test contrasts not evident in the main model of accuracy, follow-up models were built on targeted subsets of the data with recoded reference levels of the Group, Language, and/or Trial Type factors. These models revealed that on Thai "change" trials, MEBs were significantly more likely to be accurate compared to EIBs, but not compared to EMBs [$\beta = 0.339$, $z = 1.574$, $p = 0.116$]. On Thai "no change" trials, MEBs were more likely to be accurate than EIBs [$\beta = 0.749$, $z = 2.382$, $p = 0.017$] and EMBs [$\beta = 1.094$, $z = 3.520$, $p < 0.001$], but EMBs were not more likely to be accurate than EIBs [$\beta = -0.344$, $z = -1.135$, $p = 0.256$]. On Yoruba "change" trials, a different pattern of between-group differences was observed than on Thai: MEBs were again more likely to be accurate than EIBs [$\beta = 0.878$, $z = 4.969$, $p < 0.0001$] and EMBs [$\beta = 0.497$, $z = 2.826$, $p = 0.005$], but here EMBs were also more likely to be accurate than EIBs [$\beta = 0.381$, $z = 2.109$, $p = 0.035$]. On Yoruba "no change" trials, only MEBs were more likely to be accurate than EIBs [$\beta = 1.758$, $z = 5.825$, $p < 0.0001$]; EMBs were not [$\beta = 0.174$, $z = 0.621$, $p = 0.535$], and they were also less likely to be accurate than MEBs [$\beta = -1.584$, $z = -5.251$, $p < 0.0001$]. In short, the Group × Language interaction reflected the fact that group differences in accuracy varied by language, with EMBs showing an advantage over EIBs on "change" trials in Yoruba but not in Thai (counter to prediction P3).

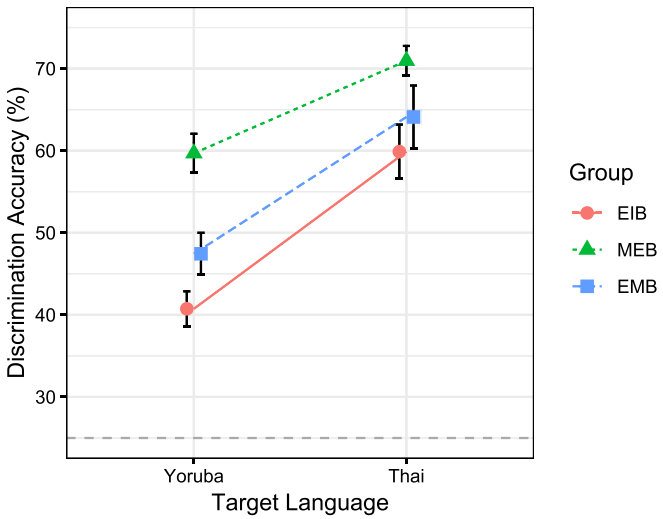


FIG. 3. (Color online) Overall accuracy in discriminating nonnative tonal contrasts in Experiment 1, by target language (in order of increasing typological similarity to Mandarin) and group. The L1 English-L2 intonational (EIB), L1 Mandarin-L2 English (MEB), and L1 English-L2 Mandarin (EMB) groups are represented, respectively, in circles, triangles, and squares. Error bars mark ±1 standard error (SE) over participants; the dotted horizontal line marks chance-level performance over both "change" and "no change" trials (= 25%).

Apart from accuracy, discrimination performance was also assessed in terms of response time (RT). In particular, RTs were examined due to the possibility of group disparities in accuracy arising from systematic differences in response speed. To investigate this possibility, log-transformed RTs for accurate responses were analyzed with linear mixed-effects regression, using `lmer()` in the `lmerTest` package (Kuznetsova *et al.*, 2019) and excluding outlier RTs greater than 2.5 SD from the participant's mean (3.1% of the data). The structure of the main model of RT was the same as in the main model of discrimination accuracy. An ANOVA on this model showed significant main effects of Group [$\chi^2(2) = 11.253$, $p = 0.004$] and Order [$\chi^2(1) = 70.740$, $p < 0.0001$], but not of Language [$\chi^2(1)=0.513$, $p = 0.474$] or Trial Type [$\chi^2(1)=1.960$, $p = 0.162$]; in addition, the Group × Language interaction was significant [$\chi^2(2) = 23.789$, $p < 0.0001$]. The Group × Trial Type [$\chi^2(2) = 5.920$, $p = 0.052$], Language × Trial Type [$\chi^2(1) = 3.321$, $p = 0.068$], and Group × Language × Trial Type [$\chi^2(2) = 5.305$, $p = 0.070$] interactions were all marginal. Table V summarizes the fixed-effect coefficients in this model.

The variation of (log) RT across groups and languages is depicted in Fig. 4. As shown in Fig. 4, the RT data generally mirrored the accuracy data: RTs were fastest for MEBs ($M = -0.407$), followed by EMBs ($M = -0.171$) and then EIBs ($M = -0.093$). However, the effect of Language was not consistent across groups, with EIBs and EMBs showing considerably slower RTs on Yoruba than Thai (Yoruba: $M_{EIB} = 0.053$, $M_{EMB} = -0.038$; Thai: $M_{EIB} = -0.124$, $M_{EMB} = -0.203$) but MEBs showing virtually no difference between the two languages ($M = -0.400$ on Yoruba vs -0.409 on Thai). The effect of trial type was also not consistent, with MEBs showing faster

TABLE V. Fixed-effect terms in the main linear mixed-effects model of response time in Experiment 1 (tone discrimination). Significance codes: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Predictor	β	SE	t	p
(Intercept)	0.012	0.073	0.160	0.873
Group: MEB	−0.261	0.097	−2.677	0.010*
Group: EMB	−0.078	0.100	−0.784	0.437
Language: Yoruba	0.049	0.046	1.053	0.293
Trial Type: no change	−0.018	0.063	−0.281	0.779
Order	−0.001	0.0001	−8.411	<0.0001***
Group: MEB × Language: Yoruba	−0.207	0.043	−4.762	<0.0001***
Group: EMB × Language: Yoruba	−0.034	0.046	−0.740	0.459
Group: MEB × Trial Type: no change	−0.184	0.057	−3.226	0.001**
Group: EMB × Trial Type: no change	−0.069	0.062	−1.114	0.265
Language: Yoruba × Trial Type: no change	0.025	0.116	0.217	0.828
Group: MEB × Language: Yoruba × Trial Type: no change	0.242	0.108	2.233	0.026*
Group: EMB × Language: Yoruba × Trial Type: no change	0.096	0.119	0.809	0.419

RTs on "no change" trials ($M = -0.469$, cf. −0.399 on "change"), EMBs showing little difference between the two trial types ($M = -0.180$ on "no change," cf. −0.171 on "change"), and EIBs showing slower RTs on "no change" trials ($M = -0.077$, cf. −0.095 on "change"). These patterns were reflected in the main effect of Group and non-significant effects of Language and Trial Type from above. Apart from a general tendency for RTs to get faster over the course of the experiment [$\beta = -0.001$, $t = -8.411$, $p < 0.0001$], the coefficients of the main model (Table V) showed that, on "change" trials in Thai, MEBs responded significantly faster than EIBs [$\beta = -0.261$, $t = -2.677$, $p = 0.010$], whereas EMBs did not [$\beta = -0.078$, $t = -0.784$, $p = 0.437$]. Additionally, the interaction coefficients indicated that the small slowdown observed on Yoruba "change" trials for EIBs was significantly reduced—in fact, reversed—for MEBs [$\beta = -0.207$, $t = -4.762$, $p < 0.0001$], although this effect was basically nullified on Yoruba "no change" trials [$\beta = 0.242$, $t = 2.233$, $p = 0.026$]. Furthermore, the small speed-up observed on Thai "no change" trials for EIBs was significantly larger for MEBs [$\beta = -0.184$, $t = -3.226$, $p = 0.001$].

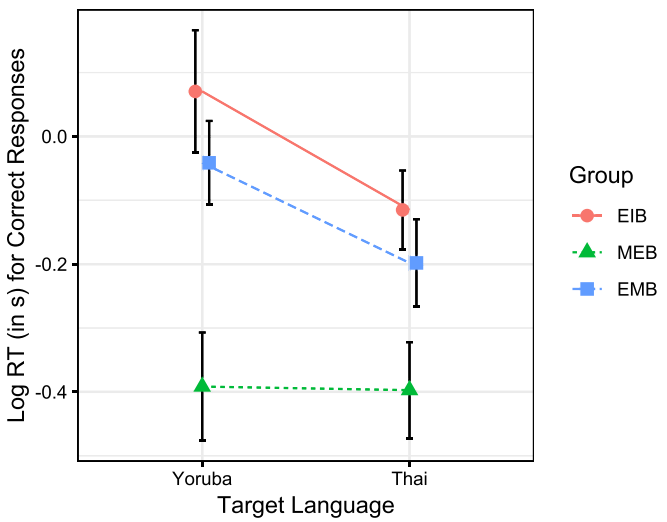


FIG. 4. (Color online) Overall response time (log) for correct responses in Experiment 1, by target language (in order of increasing typological similarity to Mandarin) and group. The L1 English-L2 intonational (EIB), L1 Mandarin-L2 English (MEB), and L1 English-L2 Mandarin (EMB) groups are represented, respectively, in circles, triangles, and squares. Error bars mark ±1 *SE* over participants.

To test contrasts not evident in the main model of RT, follow-up models were built on targeted subsets of the data with recoded reference levels of the Group, Language, and/or Trial Type factors. These models revealed that on Thai "change" trials, MEBs were significantly faster than EIBs, but only marginally faster than EMBs [$\beta = -0.183$, $t = -1.874$, $p = 0.066$]. On Thai "no change" trials, MEBs were faster than EIBs [$\beta = -0.445$, $t = -4.045$, $p < 0.001$] and EMBs [$\beta = -0.298$, $t = -2.685$, $p = 0.009$], but EMBs were not significantly faster than EIBs [$\beta = -0.147$, $t = -1.286$, $p = 0.202$]. On Yoruba "change" trials, a different pattern of between-group differences was observed than on Thai: MEBs were again faster than EIBs [$\beta = -0.475$, $t = -4.014$, $p < 0.001$], but here they were also significantly faster than EMBs [$\beta = -0.349$, $t = -2.973$, $p = 0.004$], who were not significantly faster than EIBs [$\beta = -0.126$, $t = -1.031$, $p = 0.307$]. On Yoruba "no change" trials, MEBs were again faster than both EIBs [$\beta = -0.406$, $t = -2.922$, $p = 0.004$] and EMBs [$\beta = -0.347$, $t = -2.533$, $p = 0.013$], while EMBs were not significantly faster than EIBs [$\beta = -0.059$, $t = -0.401$, $p = 0.689$]. In short, the Group × Language interaction for RTs arose because group differences varied by language, largely due to the fact that only MEBs' RTs, clearly the fastest of all groups, were unaffected by language. Thus, RTs provided converging evidence of MEBs' perceptual advantage over the other two groups. The pattern of group differences in RT, as well as accuracy, is summarized in Table VI.

The final part of the analysis of tone discrimination performance focused on accuracy across different tonal contrasts, which varied considerably both in Yoruba and in Thai. As shown in Fig. 5, all listener groups, including MEBs, showed comparatively low accuracy on the Yoruba Y2-Y3 (mid vs high) and Thai H1-H2 (mid vs low), H2-H3 (low vs falling), and H4-H5 (high vs rising) contrasts. To analyze this pattern statistically, separate logistic mixed-effects models were built for each group on each target L3, with the same random-effects structure as the model in

TABLE VI. Summary of between-group differences in discrimination accuracy and response time in Experiment 1, by target language and trial type (≫ "significantly more accurate/fast than"; > "non-significantly more accurate/fast than").

Target L3	Trial type	Accuracy	Response time
Thai	change	MEB > EMB > EIB	MEB > EMB > EIB
Thai	no change	MEB ≫ {EIB > EMB}	MEB ≫ {EMB > EIB}
Yoruba	change	MEB ≫ EMB ≫ EIB	MEB ≫ {EMB > EIB}
Yoruba	no change	MEB ≫ {EMB > EIB}	MEB ≫ {EMB > EIB}

Table IV but a deviation-coded (meaning the contrast estimate is against the grand mean rather than a reference level) fixed effect for Contrast. These models confirmed that, for all groups, the likelihood of accuracy was significantly lower than average for Y2-Y3 [βs < -0.437, zs < -2.361, ps < 0.05], H1-H2 [βs < -1.125, zs < -5.898, ps < 0.0001], and H4-H5 [βs < -2.489, zs < -12.095, ps < 0.0001]. However, H2-H3 showed a likelihood of accuracy significantly lower than average only for MEBs [$\beta = -1.612$, $t = -8.373$, $p < 0.0001$] and not for EIBs or EMBs [$|\beta|$s < 0.158, $|z|$s < 0.822, ps > 0.05]. At the same time, the likelihood of accuracy was significantly higher than average for other contrasts, such as Y1-Y3 for EMBs and MEBs [βs > 0.501, zs > 2.727, ps < 0.01] and H3-H5 for all groups [βs > 0.808, zs > 4.841, ps < 0.0001].

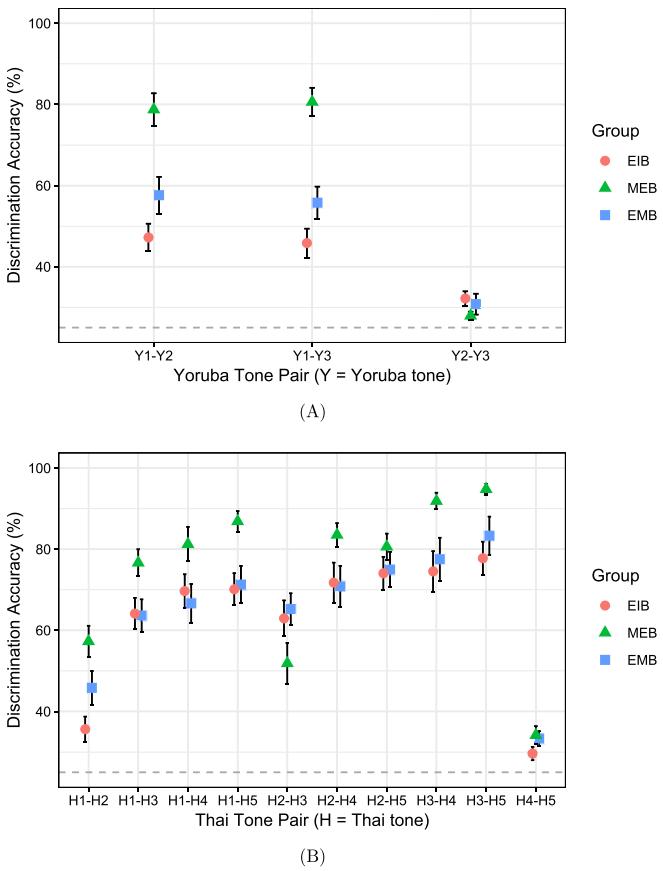


FIG. 5. (Color online) Accuracy in discriminating (A) Yoruba and (B) Thai tonal contrasts in Experiment 1, by tone pair and group. The L1 English-L2 intonational (EIB), L1 Mandarin-L2 English (MEB), and L1 English-L2 Mandarin (EMB) groups are represented, respectively, in circles, triangles, and squares. Error bars mark ±1 *SE* over participants; the dotted horizontal line marks chance-level performance over "change" trials assuming that some change was successfully detected (= 33%).

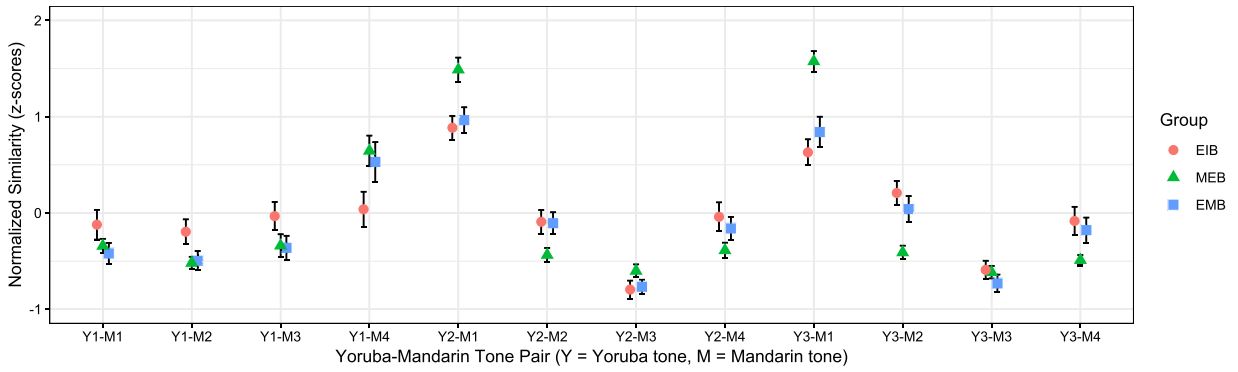
Given previous findings of L2 Mandarin transfer that was specific to L3 tone pairs contrasting in pitch direction (Qin and Jongman, 2016), variation in accuracy across different tonal contrasts was also analyzed in terms of a dichotomy between "height pairs" (i.e., tonal contrasts based primarily on pitch height) and "direction pairs" (i.e., tonal contrasts based primarily on pitch direction), in order to see whether between-group differences were specific to certain contrast types. For this analysis, the contrasts classified as clear "height pairs" were Y2-Y3 and H1-H2 (because the general directionality, if not the slope, of any pitch change is the same for the two tones), while those classified as clear "direction pairs" were H1-H4 and H2-H5 (because a substantial portion, i.e., 50% or more, of the tone contour occurs at around the same pitch height for the two tones; note that, consistent with its classification as a register tone language, there are no clear "direction pairs" in Yoruba). This analysis did not reveal any systematic effect of the "height pair" vs "direction pair" dichotomy. More specifically, the perceptual advantage that EMBs, as well as MEBs, showed over EIBs was not targeted toward "direction pairs." For example, EMBs did not show an advantage over EIBs on the "direction pairs" H1-H4 and H2-H5, while they did show an advantage—in fact, the clearest advantage—on the "height pair" H1-H2 (Fig. 5).

In short, the results of the by-contrast analyses were consistent with our prediction regarding the effect of acoustic similarity on relative discriminability (P5), showing a similar pattern of variation across contrasts in all groups, including EIBs. Notably, however, whereas MEBs had the highest accuracy of all groups on nearly every contrast, they showed comparatively low accuracy on the Thai H2-H3 (low vs falling) contrast, actually the lowest of the three groups. To explore this between-group variation across contrasts further, we examined the perceptual similarity data gathered in Experiment 2.

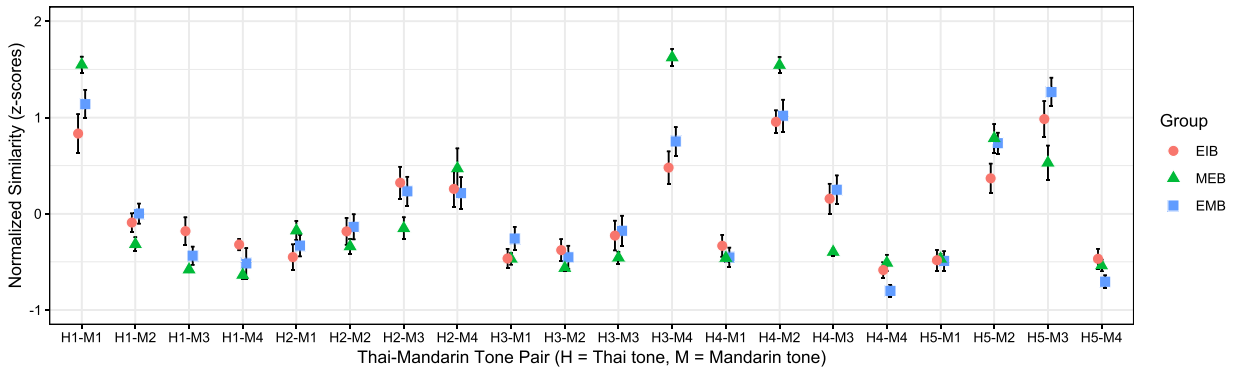
B. Experiment 2: Crosslinguistic perceptual similarity

Recall that the purpose of Experiment 2 was to see whether, in line with the Perceptual Assimilation Model, differences in crosslinguistic perceptual similarity between L3 and known (Mandarin) tones could account for variation in the discriminability of L3 tonal contrasts. To control for variation across listeners in use of the rating scale and to address scale compression and skew (see Schütze and Sprouse, 2014), all similarity ratings were normalized by listener, relative to the full set of that listener's tone similarity ratings. Mean normalized similarity ratings for each L3-Mandarin tone pair are shown in Fig. 6. Note that EIBs did not actually know Mandarin, so their similarity ratings can be interpreted as primarily reflecting the acoustic similarity of the paired tones.

As with accuracy, tone similarity ratings showed considerable variation across tone pairs, with a high degree of perceived similarity for certain tone pairs (i.e., normalized similarity ratings significantly higher than zero; all ps < 0.0001 in one-tailed t-tests). In the case of Yoruba, Y2 (mid) and Y3 (high)



(A)



(B)

FIG. 6. (Color online) Perceptual similarity between (A) Yoruba or (B) Thai tones and Mandarin tones in Experiment 2, by tone pair and group. The *x* axis shows tone pairs in the order Yoruba or Thai, then Mandarin. The L1 English-L2 intonational (EIB), L1 Mandarin-L2 English (MEB), and L1 English-L2 Mandarin (EMB) groups are represented, respectively, in circles, triangles, and squares. Error bars mark ±1 *SE* over participants.

were each perceived as very similar to Mandarin M1 (high level), with mean normalized ratings of 1.1 and 1.0, respectively; this was the case for all groups, especially MEBs (see Fig. 6). Certain Thai-Mandarin tone pairs were also perceived as similar: H4 (high) and H5 (rising) were both rated as similar to M2 (mid rising), although H4 was rated as more similar to M2 ($M_{sim} = 1.2$) than was H5 ($M_{sim} = 0.6$). These results were consistent with patterns observed in Experiment 1, which showed that two of the four least discriminable L3 tonal contrasts (for all listeners, including Mandarin-speaking bilinguals) were, in fact, Y2-Y3 and H4-H5. On the other hand, the low discriminability of the H1-H2 contrast seen in Experiment 1 was not reflected in similarity data, as these two tones were perceived as most similar to different Mandarin tones—H1 as most similar to M1, and H2 as most similar to M3 and/or M4.

Additional tone pairs were rated as perceptually similar as well. Both H2 (low) and H3 (falling) were rated as similar to M4 (high falling), although the groups differed with respect to how clearly they identified M4 as the target of crosslinguistic perceptual assimilation for these Thai tones. Whereas MEBs perceived H2 as clearly most similar to M4, EMBs and EIBs were split between M3 and M4, giving similarity ratings for the two that were virtually identical [see Fig. 6(B)]. Thus, for MEBs, who perceived both Thai tones as most similar to the same Mandarin tone but with different degrees of similarity, the H2-H3 contrast was likely to undergo Category-Goodness assimilation. By contrast, for EMBs who did not clearly perceive the Thai tones as most similar to the same Mandarin tone, the H2-H3 contrast was not likely to undergo even Category-Goodness assimilation (as would have been the case necessarily for EIBs who had no knowledge of Mandarin or of tones). As such, the Perceptual Assimilation Model predicts that discrimination of this L3 contrast should be more difficult for MEBs than EMBs, which was in fact observed: MEBs' accuracy on H2-H3 was significantly lower than EMBs' ($M_{MEB} = 52\%$, $M_{EMB} = 65\%$; $\beta = -1.183$, $z = -2.383$, $p = 0.017$).

Other L3 tones were perceived as more different from Mandarin tones, and/or as similar to distinct Mandarin tones, and these patterns of crosslinguistic perceptual similarity were associated with greater discriminability. For example, Y1 (low) and Y3 (high) were rated by all groups as most similar to M4 (high falling) and M1 (high level), respectively, while H3 (falling) and H5 (rising) were rated by all groups as most similar to M4 and M2/M3 (mid rising, low dipping), respectively, suggesting that these tone pairs tended to undergo Two-Category assimilation for Mandarin speakers. Accordingly, the Y1-Y3 contrast and the H3-H5 contrast were both discriminated relatively well—by all groups, especially MEBs ($M_{acc} = 81\%$ for Y1-Y3, $M_{acc} = 95\%$ for H3-H5).

As for holistic perceptual similarity, listeners perceived neither Yoruba nor Thai as sounding very similar to Mandarin. The mean normalized similarity ratings based on continuous speech were, overall, −0.213 (*SD* 0.837) for the

Yoruba-Mandarin comparison and −0.116 (*SD* 1.029) for the Thai-Mandarin comparison. By group, the mean normalized similarity ratings for the Yoruba-Mandarin comparison were –0.471 (EIB), 0.012 (MEB), and −0.206 (EMB), while those for the Thai-Mandarin comparison were −0.320 (EIB), 0.260 (MEB), and −0.328 (EMB). The modest differences between the two comparisons were not significant, across groups or for any individual group [paired-sample $|t|$s < 1.261, ps > 0.05].

V. DISCUSSION

A. Synthesis of the findings

The goal of the present study was to examine how nonnative tonal contrasts are perceived by sequential bilinguals, with a view toward informing research on L3 acquisition of tone and other understudied suprasegmental features. Comparing L1 Mandarin-L2 English bilinguals (MEBs), L1 English-L2 Mandarin bilinguals (EMBs), and L1 English-speaking bilinguals with no tonal experience (EIBs), the current study found that MEBs had a robust perceptual advantage over EMBs and EIBs in discriminating tones from an unfamiliar L3, outperforming them in both accuracy and response speed and on two different L3s. However, the benefits of previous tonal experience were much smaller in EMBs. Although numerical differences between EMBs and EIBs generally favored EMBs, EMBs showed a statistically significant advantage over EIBs only in terms of accuracy on Yoruba "change" trials (Table VI). Together these findings support our first two predictions (P1, P2) by suggesting that prior tonal experience facilitates L3 tone perception, with the benefit derived from this experience being greater when coming from the L1 than the L2. However, our conclusion concerning the acquisition order effect should be considered tentative, since it is not clear that the MEBs and EMBs in this study had similar Mandarin perceptual abilities available to transfer; data on EMBs' L2 Mandarin perception are thus needed to confirm this interpretation of the different benefits of tonal experience shown by MEBs and EMBs.

As for our third prediction (P3), although we found no evidence of a difference between the two target L3s in terms of holistic perceived similarity with Mandarin, there was still a difference in typological similarity, Thai being more similar than Yoruba to Mandarin (see Sec. II B); nevertheless, this difference did not correspond to greater benefits of prior tonal experience on Thai. Contrary to expectation, greater benefits of prior tonal experience were observed on Yoruba: here, MEBs showed a larger advantage in accuracy over EMBs and EIBs, while EMBs showed a more robust advantage over EIBs (Fig. 3). The main effect of language—higher levels of accuracy on Thai than Yoruba, across groups—was not predicted by P3. Crucially, however, this effect was also evident in EIBs (i.e., tonally naive listeners), suggesting that the language effect was not due to similarity with Mandarin. Rather, Yoruba tones appear to have been objectively more difficult to discriminate than Thai tones, leading to greater apparent benefits of prior tonal experience on the relatively more challenging Yoruba contrasts.

Finally, comparisons of discrimination outcomes against crosslinguistic perceptual similarity data provided limited support for our fourth prediction (P4) related to variation across contrasts. As expected, certain L3 tonal contrasts were significantly less discriminable than others, and the lower discriminability of an L3 contrast was often reflected in perceived similarity of the two L3 tones to the same Mandarin tone. However, this was not the case for the less discriminable Thai H1-H2 contrast, suggesting that perceptual assimilation to previously-acquired categories provides one, but not the only, explanation for the difficulty of certain nonnative contrasts. More generally, we interpret the correlation between lower discriminability and convergent patterns of crosslinguistic perceptual similarity cautiously because it was not only Mandarin speakers, but also non-Mandarin speakers (EIBs), who showed the relevant patterns of perceptual similarity. The fact that non-Mandarin speakers could not assimilate L3 tones to Mandarin, yet also showed decreased levels of discrimination for all of the same L3 contrasts that Mandarin speakers did, indicates that there must be other factors leading to lower discriminability besides disadvantageous patterns of perceptual assimilation. In particular, the basic acoustic phonetic similarity of certain tones (e.g., Y2-Y3, H1-H2) appeared to make them less discriminable for all listeners, including EIBs, consistent with our fifth prediction (P5).

B. Implications for models of L2 and L3 acquisition

Notably, the results of this study are not fully predicted by any of the main L3 acquisition models discussed above (Cumulative Enhancement Model, L2 Status Factor Model, TPM). In regard to L3 tone perception, our findings provide suggestive evidence that prior tonal experience does not transfer specifically from the L2, in a cumulative fashion from the L1 and/or L2, or on the basis of typological similarity between known and target languages. Further, none of these models predicts transfer primarily from the L1, as was found here for tone perception in two different L3s. However, this pattern is in line with previous findings of L1 transfer in L3 acquisition (Hermas, 2014; Lozano, 2003) and can be explained in terms of the ASP framework (Strange, 2011). Under the ASP view, when listeners detect phonologically relevant information in the acoustic signal, this activates the selective perception routines developed for a previously-acquired language (e.g., L1), which guide them toward processing the signal efficiently in service of identifying contrastive categories. Due to the dominant influence of L1 selective perception routines (cf. Kuhl, 2000), therefore, the superior performance of L1 Mandarin speakers (MEBs) can be attributed to the necessarily high attunement to pitch in their L1 selective perception routines—in particular, pitch information on the short timescale of lexical tones.

Crucially, however, L1 speakers of Mandarin, as compared to L2 speakers, showed a much clearer benefit of prior tonal experience in L3 tone perception, suggesting a privileged role for L1 (i.e., early) exposure to acoustic cues as linguistically informative. Such a powerful effect of L1 experience has also been found in other cases of bilingual speech perception, such as the perception of nonnative stop laryngeal contrasts (McKelvie-Sebileau and Davis, 2014). In regard to tone, the current results support the view that the

timing of linguistic experience with tones has significant consequences for how phonetic cues crucial for identifying tones are perceived in a new language. In particular, listeners exposed to tones later in life may process, represent, and/or access pitch variation on the timescale of tones differently—and perhaps less efficiently—than listeners with early exposure to tones. For example, Mandarin speakers have been found to attend to tone contours subconsciously and to employ knowledge of phonological rules governing tone change in perceiving Mandarin tones, whereas non-tonal language speakers tend to rely more heavily on pitch height (Huang and Johnson, 2010). More generally, due to the nature of tone languages, tone-language speakers are constantly engaged in processing pitch variation at the lexical level. Such linguistic experience, as shown by Chang *et al.* (2017), may be transferred beneficially to assist initial perception of tonal contrasts in another language.

Whereas ASP does predict L1 tonal experience to transfer to L3 tone perception, it does not explicitly predict L2 tonal experience to fail to transfer. On the contrary, because L2 learners are understood to develop selective perception routines for the L2 (i.e., not merely to transfer L1 selective perception routines), a logically possible outcome under ASP is for L2 learners of a tone language to transfer their L2 tonal experience to L3 tone perception. This is why proficient EMBs, who were assumed to have developed L2 selective perception routines for perceiving Mandarin tones, were predicted to outperform EIBs in L3 tone perception (see Table II). Indeed, EMBs outperformed EIBs in terms of accuracy on Yoruba "change" trials; however, this was the only place where their advantage over EIBs was statistically significant. EMBs' relatively weak advantage, in comparison to MEBs' robust advantage, therefore points to either (or both) of the following conclusions: (1) L2 selective perception routines generally transfer more weakly to L3 perception than do L1 selective perception routines, or (2) the specific EMBs tested in this study had L2 selective perception routines that were much less effective for perceiving tone than the L1 selective perception routines of MEBs (i.e., despite the importance of tone in Mandarin, EMBs were still not particularly attuned to pitch).

The findings of Qin and Jongman (2016), who examined L2 learners of Mandarin with less Mandarin exposure and lower Mandarin proficiency than the EMBs in the current study, provide additional data relevant to both conclusions. On the one hand, their results cast doubt on the second conclusion because their L2 Mandarin learners clearly had strong tone perception abilities, outperforming L1 Mandarin listeners at Cantonese tone discrimination with accuracy levels of over 90%. On the other hand, their results are not entirely consistent with the first conclusion either because their L2 Mandarin learners' sensitivity to pitch direction did apparently transfer to L3 tone perception, albeit in a simpler task. The difference in task demands between Qin and Jongman (2016) and the current study is worth noting, given the influential role of task demands in ASP (Strange, 2011). For example, it is possible that the higher task demands in the current study led to facilitative transfer of previous perceptual experience, and/or the blocking of non-facilitative transfer, occurring less effectively compared to what was observed in Qin and Jongman (2016); this would obviate the need to posit a general acquisition order effect (i.e., L1-L2 difference) in transfer of selective perception routines, thereby implicating insufficiently high L2 phonological proficiency as the reason for EMBs' weak advantage over EIBs. Unfortunately, however, EMBs' Mandarin (L2) phonological proficiency (in particular, their Mandarin tone perception) was not tested directly—a design limitation of this study that should be avoided in future L3 acquisition research, which should ideally include measures in all of the learners' languages (i.e., L1, L2, and L3). In the end, the fact remains that both of the above conclusions (which are not mutually exclusive) are consistent with the information available in this study, so we remain agnostic as to the explanation for EMBs' relatively weak advantage over EIBs.

In addition to the implications for L3 acquisition models, the current findings also have implications for the Perceptual Assimilation Model, which has generally been applied to L2 (as opposed to L3) perception. Overall, the results concerning individual L3 contrasts were consistent with the core prediction of the Perceptual Assimilation Model: higher discriminability tended to correspond to a Two-Category pattern of perceptual assimilation to Mandarin tones, while lower discriminability tended to correspond to a Single-Category or Category-Goodness pattern (Figs. 5 and 6). However, this correspondence did not hold for all L3 contrasts. In particular, the Thai H1-H2 contrast showed low discriminability in all groups despite a Two-Category pattern of perceptual assimilation. Furthermore, although perceptual assimilation to Mandarin categories was only possible for the groups that actually knew Mandarin (MEBs, EMBs), the group that did not know Mandarin (EIBs) nevertheless showed a highly similar pattern of variation in discriminability across L3 contrasts, suggesting that much of this variation may be attributable to (pre-categorical) acoustic phonetic similarity, as opposed to category-level influence per se. In short, the results of this study, while largely providing empirical validation for the category-based approach of the Perceptual Assimilation Model, also reveal the limitations of such an approach to predicting patterns of L3 speech perception.

VI. CONCLUSION

Although the current findings are broadly consistent with an "L1 Status Factor" in L3 perception, it should be noted that they are based on comparisons of specific bilingual language backgrounds, and different results could be found with bilinguals of other backgrounds. In particular, even though the L2 Mandarin speakers tested in this study were, on average, relatively proficient (according to both self-ratings and proficiency test scores), they represented a limited part of the continuum of Mandarin proficiency and, for the most part, instructed (as opposed to naturalistic) L2 learners (cf. Cabrelli Amaro and Wrembel, 2016). Furthermore, their phonological proficiency in the L1/L2 was not tested specifically, which prevents us from confirming that they had developed reliable selective perception

routines for perceiving Mandarin tones. Thus, it is possible that L2 Mandarin speakers with higher overall Mandarin proficiency, clearly advanced phonological proficiency in Mandarin, and/or extensive naturalistic experience learning Mandarin might end up showing a more robust advantage over non-tonal bilinguals than was observed in this study.

Despite these limitations, this study sheds new light on the nature of language transfer in L3 perceptual development, with implications for theoretical models of L3 acquisition and future research in this area. To improve our understanding of transfer in L3 phonetic and phonological acquisition, it will be crucial to replicate the current findings on transfer (or the lack thereof) of prior tonal experience with bilinguals representing different language profiles, which may involve investigating Mandarin learners from other L1 backgrounds (e.g., French, Korean), speakers of other tone languages (e.g., Vietnamese), speakers of languages exemplifying other prosodic types (e.g., "pitch accent" languages such as Japanese), and speakers representing different combinations and trajectories of L1 and L2 proficiency. In addition, direct measurement of L2 phonetic and phonological abilities, which should be the standard in future research on L3 phonetics and phonology, will allow for more nuanced investigations of the links between the L1, L2, and L3 in incipient multilingualism. These research avenues, in addition to the examination of other modalities (e.g., production) and the many other features that make up a language's phonology, are sure to provide valuable insights into the role of bilingual knowledge in shaping the course of L3 development.

ACKNOWLEDGMENTS

The authors thank the Boston University Center for the Humanities for financial support, as well as the audience at BUCLD 42, especially Jennifer Cabrelli, Jason Rothman, and Tyler Perrachione, and several anonymous reviewers for helpful feedback.

[1]Although the conventional abbreviation for "tone" in the tone literature is "T," we will abbreviate with the first non-T letter of the language name so as to prevent confusion between tones from different languages that are numbered the same.

[2]Note that if, instead of considering the particular shape of the pitch contour, the initial part of the pitch contour were ignored and the overall directionality of pitch change taken to be the crucial variable for comparing contours, two contrasts (H2-H3, H4-H5) could be interpreted as register contrasts in Thai. However, even in this case, Thai's tone system would less "register-like" than Yoruba's, with 20% (2/10) register contrasts compared to 33% (1/3) in Yoruba.

[3]Although neither Thai (a member of the Kra-Dai language family) nor Yoruba (a member of the Niger-Congo language family) is genetically related to Mandarin (a member of the Sino-Tibetan language family), it is reasonable to wonder whether, due to geographic proximity, Thai might have borrowed some lexical items from Mandarin, resulting in greater lexical overlap between Thai and Mandarin as compared to Yoruba and Mandarin. In short, the only lexical overlap between Thai and any dialect of Chinese that we are aware of is with the Teochew dialect, which is mutually unintelligible with Mandarin. Nevertheless, to check our assumption regarding (lack of) lexical overlap, we had a native Mandarin speaker, a trained linguist with no knowledge of Thai or Yoruba, listen to continuous speech passages we recorded in both languages, telling her in advance that both comprised excerpts from Aesop's fable "The North Wind and the Sun" (see Sec. III C 2) and asking her specifically to listen for words that seemed similar to lexical items of Mandarin. Even with this advance knowledge of the semantic content she was going to hear, she noticed no items, in either passage, that sounded like any lexical item of Mandarin. Consequently, we believe it is reasonable to assume that the two L3s show equal (namely, zero) lexical overlap with Mandarin.

[4]Although tones are often referred to here with their pitch-based descriptors along with their numerical labels, this is strictly for ease of exposition and should not be construed as implying that pitch properties are the only relevant cues for discriminating tones or for perceiving tones as similar.

[5]The one MEB participant who reported knowledge of Cantonese said that she was exposed to Cantonese in early childhood only and, at the time of testing, could no longer speak Cantonese. Her data were not exceptional within the MEB group: her overall discrimination accuracy was 40% for Yoruba (cf. range of 33%–71% among MEBs) and 65% for Thai (cf. range of 54%–86% among MEBs). Excluding this participant from the MEB group did not affect the results, so her data are included in all analyses.

[6]For more information, see https://osf.io/3fb8p/.

[7]See the full questionnaire at https://osf.io/jtpzh/.

[8]The raw demographic data are available at https://osf.io/3fb8p/.

[9]See supplementary material at https://doi.org/10.1121/1.5120522 for a table summarizing language background information on all participant groups.

[10]For more information, see https://osf.io/3fb8p/.

[11]An anonymous reviewer wondered whether the higher accuracies on Thai could be due to the larger percentage of "change" trials in the Thai condition as compared to the Yoruba condition (92% in Thai vs 86% in Yoruba; see Sec. III C 1) leading to a relatively stronger bias to identify a change in the Thai condition. Although we cannot rule out this possibility, we consider this unlikely, because there is no reason to expect such a response bias to be isolated to "change" trials, yet we see no evidence of it in "no change" trials. To be specific, a stronger bias toward identifying a change in Thai should lead to a greater incidence of "false alarms" (i.e., incorrect identification of a change) on "no change" trials in Thai. Crucially, however, the data evince the opposite trend: whereas false alarms account for 10.9% of the errors (and 5.5% of all responses) on Yoruba, they account for 4.8% of the errors (and 1.6% of all responses) on Thai.

Abramson, A. S. (**1962**). "The vowels and tones of standard Thai: Acoustical measurements and experiments," Int. J. Am. Linguist. **28**(2), 1–155.

Abramson, A. S. (**1976**). "Thai tones as a reference system," in *Tai Linguistics in Honor of Fang-Kuei Li*, edited by T. W. Gething, J. G. Harris, and P. Kullavanijaya (Chulalonkorn University Press, Bangkok, Thailand), pp. 1–12.

Agwuele, A. H. (**2005**). "'Yorubaisms' in African American "speech" patterns," in *The Yoruba Diaspora in the Atlantic World*, edited by T. Falola and M. D. Childs (Indiana University Press, Bloomington, IN), pp. 325–345.

Bamgbose, A. (**1967**). *A Short Yoruba Grammar* (Heinemann Educational Books, Ibadan, Nigeria).

Bardel, C., and Falk, Y. (**2007**). "The role of the second language in third language acquisition: The case of Germanic syntax," Second Lang. Res. **23**(4), 459–484.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Grothendieck, G., Scheipl, F., Green, P., and Fox, J. (**2019**). "Package 'lme4': Linear mixed-effects models using 'Eigen' and S4," R package (version 1.1-21) [computer program], https://cran.r-project.org/web/packages/lme4/ (Last viewed July 12, 2019).

Bent, T., Bradlow, A. R., and Wright, B. A. (**2006**). "The influence of linguistic experience on the cognitive processing of pitch in speech and nonspeech sounds," J. Exp. Psychol. Hum. Percept. Perform. **32**(1), 97–103.

Berkes, É., and Flynn, S. (**2012**). "Further evidence in support of the Cumulative-Enhancement Model: CP structure development," in *Third Language Acquisition in Adulthood*, edited by J. Cabrelli Amaro, S. Flynn, and J. Rothman (John Benjamins Publishing, Amsterdam, The Netherlands), pp. 143–164.

Best, C. T. (**1994**). "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 J. C. Goodman and H. C. Nusbaum (MIT Press, Cambridge, MA), pp. 167–224.

Best, C. T., and Strange, W. (**1992**). "Effects of phonological and phonetic factors on cross-language perception of approximants," J. Phon. **20**(3), 305–330.

Best, C. T., and Tyler, M. D. (**2007**). "Nonnative and second-language speech perception: Commonalities and complementarities," in *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*, edited by O.-S. Bohn and M. J. Munro (John Benjamins Publishing, Amsterdam, The Netherlands), pp. 13–34.
Boersma, P., and Weenink, D. (**2016**). "Praat: Doing phonetics by computer, (version 6.0.19) [computer program]," http://www.praat.org (Last viewed July 12, 2019).
Brysbaert, M. (**2013**). "LEXTALE_FR: A fast, free, and efficient test to measure language proficiency in French," Psychol. Belgica **53**(1), 23–37.
Burnham, D., Francis, E., Webster, D., Luksaneeyanawin, S., Attapaiboon, C., Lacerda, F., and Keller, P. (**1996**). "Perception of lexical tone across languages: Evidence for a linguistic mode of processing," in *Proceedings ICSLP 96: Fourth International Conference on Spoken Language Processing*, Citation Delaware, October 3–6, New Castle, DE, pp. 2514–2517.
Burnham, D., and Jones, C. (**2002**). "Categorical perception of lexical tone by tonal and non-tonal language speakers," in *Proceedings of the Ninth Australian International Conference on Speech Science and Technology*, December 3–5, Melbourne, Australia, pp. 515–520.
Cabrelli Amaro, J. (**2012**). "L3 phonology: An understudied domain," in *Third Language Acquisition in Adulthood*, edited by J. Cabrelli Amaro, S. Flynn, and J. Rothman (John Benjamins Publishing, Amsterdam, The Netherlands), pp. 33–60.
Cabrelli Amaro, J. (**2013**). "Methodological issues in L3 phonology," Stud. Hispanic Lusophone Lingusit. **6**(1), 101–117.
Cabrelli Amaro, J. (**2017**). "Testing the Phonological Permeability Hypothesis: L3 phonological effects on L1 versus L2 systems," Int. J. Bilingual. **21**(6), 698–717.
Cabrelli Amaro, J., and Rothman, J. (**2010**). "On L3 acquisition and phonological permeability: A new test case for debates on the mental representation of non-native phonological systems," Int. Rev. Appl. Linguist. Lang. Teach. **48**(2–3), 275–296.
Cabrelli Amaro, J., and Wrembel, M. (**2016**). "Investigating the acquisition of phonology in a third language—A state of the science and an outlook for the future," Int. J. Multilingual. **13**(4), 395–409.
Carlson, M. T., Goldrick, M., Blasingame, M., and Fink, A. (**2016**). "Navigating conflicting phonotactic constraints in bilingual speech perception," Bilingual. Lang. Cogn. **19**(5), 939–954.
Chan, I. L., and Chang, C. B. (**2018**). "LEXTALE_CH: A quick, character-based proficiency test for Mandarin Chinese," in *Proceedings of the 42nd Annual Boston University Conference on Language Development*, November 3–5, Boston, MA, pp. 114–130.
Chang, C. B. (**2018**). "Perceptual attention as the locus of transfer to nonnative speech perception," J. Phon. **68**, 85–102.
Chang, C. B., and Yao, Y. (**2016**). "Toward an understanding of heritage prosody: Acoustic and perceptual properties of tone produced by heritage, native, and second language speakers of Mandarin," Heritage Lang. J. **13**(2), 134–160.
Chang, Y. S., Yao, Y., and Huang, B. H. (**2017**). "Effects of linguistic experience on the perception of high-variability non-native tones," J. Acoust. Soc. Am. **141**(2), EL120–EL126.
Chao, Y. R. (**1930**). "A system of 'tone-letters,'" Le Maître Phon. **45**, 24–27.
Chao, Y. R. (**1968**). *Language and Symbolic Systems* (Cambridge University Press, Cambridge, UK).
Courtenay, K. R. (**1968**). "A generative phonology of Yorùbá," Ph.D. thesis, University of California, Los Angeles, Los Angeles, CA.
Falk, Y., and Bardel, C. (**2011**). "Object pronouns in German L3 syntax: Evidence for the L2 status factor," Second Lang. Res. **27**(1), 59–82.
Flege, J. E. (**2003**). "A method for assessing the perception of vowels in a second language," in *Issues in Clinical Linguistics*, edited by E. Fava and A. Mioni (Unipress, Padova, Italy), pp. 19–43.
Flege, J. E., and MacKay, I. R. A. (**2004**). "Perceiving vowels in a second language," Stud. Second Lang. Acq. **26**(1), 1–34.
Flynn, S., Foley, C., and Vinnitskaya, I. (**2004**). "The cumulative-enhancement model for language acquisition: Comparing adults' and children's patterns of development in first, second and third language acquisition of relative clauses," Int. J. Multilingual. **1**(1), 3–16.
Foote, R. (**2009**). "Transfer in L3 acquisition: The role of typology," in *Third Language Acquisition and Universal Grammar*, edited by Y. I. Leung (Multilingual Matters, Bristol, UK), pp. 89–114.
Fox, J., Weisberg, S., Price, B., Adler, D., Bates, D., Baud-Bovy, G., Bolker, B., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Maechler, M., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W., Walker, S., Winsemius, D., Zeileis, A., and R Core Team (**2018**). "Package 'car': Companion to applied regression," R package version 3.0-2. https://cran.r-project.org/web/packages/car/ (Last viewed July 12, 2019).
Gandour, J. T., and Harshman, R. A. (**1978**). "Crosslanguage differences in tone perception: A multidimensional scaling investigation," Lang. Speech **21**(1), 1–33.
Guenther, F. H., Husain, F., Cohen, M. A., and Shinn-Cunningham, B. G. (**1999**). "Effects of categorization and discrimination training on auditory perceptual space," J. Acoust. Soc. Am. **106**(5), 2900–2912.
Guion, S. G., Flege, J. E., Akahane-Yamada, R., and Pruitt, J. C. (**2000**). "An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants," J. Acoust. Soc. Am. **107**(5), 2711–2724.
Hammarberg, B. (**2010**). "The languages of the multilingual: Some conceptual and terminological issues," Int. Rev. Appl. Linguist. Lang. Teach. **48**(2–3), 91–104.
Hammarberg, B., and Hammarberg, B. (**2009**). "Re-setting the basis of articulation in the acquisition of new languages: A third language case study," in *Processes in Third Language Acquisition*, edited by B. Hammarberg (Edinburgh University Press, Edinburgh, UK), pp. 74–85.
Hermas, A. (**2014**). "Multilingual transfer: L1 morphosyntax in L3 English," Int. J. Lang. Studies **8**(2), 1–24.
Huang, T., and Johnson, K. (**2010**). "Language specificity in speech perception: Perception of Mandarin tones by native and nonnative listeners," Phonetica **67**(4), 243–267.
Izura, C., Cuetos, F., and Brysbaert, M. (**2014**). "Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size," Psicológica **35**(1), 49–66.
James, A. L. (**1923**). "The tones of Yoruba," Bull. School Oriental Studies, Univ. Lond. **3**(1), 119–128.
Jin, F. (**2009**). "Third language acquisition of Norwegian objects: Interlanguage transfer or L1 influence?," in *Third Language Acquisition and Universal Grammar*, edited by Y. I. Leung (Multilingual Matters, Bristol, UK), pp. 144–161.
Kuhl, P. K. (**1980**). "Perceptual constancy for speech-sound categories in early infancy," in *Child Phonology*, edited by G. H. Yeni-Komshian, J. F. Kavanagh, and C. A. Ferguson (Academic Press, New York), pp. 41–66.
Kuhl, P. K. (**2000**). "A new view of language acquisition," Proc. Natl. Acad. Sci. U.S.A. **97**(22), 11850–11857.
Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (**2019**). "Package 'lmerTest': Tests in linear mixed effects models," R package version 3.1-0. https://cran.r-project.org/web/packages/lmerTest/ (Last retrieved July 12, 2019).
Lamidi, M. T. (**2003**). "The tone as a negative marker in Ìjèsà sentences," Poznań Studies Contemp. Linguist. **38**, 89–101.
Lee, Y.-S., Vakoch, D. A., and Wurm, L. H. (**1996**). "Tone perception in Cantonese and Mandarin: A cross-linguistic comparison," J. Psycholinguist. Res. **25**(5), 527–542.
Lemhöfer, K., and Broersma, M. (**2012**). "Introducing LexTALE: A quick and valid lexical test for advanced learners of English," Behav. Res. Methods **44**(2), 325–343.
Leung, I. Y. (**1998**). "Transfer between interlanguages," in *Proceedings of the 22nd Annual Boston University Conference on Language Development*, edited by A. Greenhill, M. Hughes, H. Littlefield, and H. Walsh (Cascadilla Press, Somerville, MA), pp. 477-487.
Leung, Y. I. (**2003**). "Failed features versus full transfer full access in the acquisition of a third language: Evidence from tense and agreement," in *Proceedings of the 6th Generative Approaches to Second Language Acquisition Conference*, April 26–28, Ottawa, Canada, pp. 199–207.
Lim, V. P. C., Rickard Liow, S. C., Lincoln, M., Chan, Y. H., and Onslow, M. (**2008**). "Determining language dominance in English-Mandarin bilinguals: Development of a self-report classification tool for clinical use," Appl. Psycholinguist. **29**(3), 389–412.
Llama, R., Cardoso, W., and Collins, L. (**2010**). "The influence of language distance and language status on the acquisition of L3 phonology," Int. J. Multilingual. **7**(1), 39–57.
Lozano, C. (**2003**). "Focus, pronouns and word order in the acquisition of L2 and L3 Spanish," Ph.D. thesis, University of Essex, Colchester, UK.
Maimone, L. L. (**2017**). "The role of crosslinguistic influence from L2 Spanish, type of linguistic item, and aptitude in the learning stages of L3 Portuguese forms: An exploratory study," Ph.D. thesis, Georgetown University, Washington, DC.
Mathôt, S., Schreij, D., and Theeuwes, J. (**2012**). "OpenSesame: An open-source, graphical experiment builder for the social sciences," Behav. Res. Methods **44**(2), 314–324.

McKelvie-Sebileau, P., and Davis, C. (**2014**). "Discrimination of foreign language speech contrasts by English monolinguals and French/English bilinguals," J. Acoust. Soc. Am. **135**(5), 3025–3035.
Na Ranong, S., and Leung, Y. I. (**2009**). "Null objects in L1 Thai-L2 English-L3 Chinese: An empiricist take on a theoretical problem," in *Third Language Acquisition and Universal Grammar*, edited by Y. I. Leung (Multilingual Matters, Bristol, UK), pp. 162–191.
Neuser, H. (**2017**). "Source language of lexical transfer in multilingual learners: A mixed methods approach," Ph.D. thesis, Stockholm University, Stockholm, Sweden.
Ọdẹlóbí, Ọ. À. (**2008**). "Recognition of tones in Yorùbá speech: Experiments with artificial neural networks," in *Speech, Audio, Image and Biomedical Signal Processing using Neural Networks*, edited by B. Prasad and S. R. M. Prasanna (Springer, Berlin, Germany), pp. 23–47.
Onishi, H. (**2016**). "The effects of L2 experience on L3 perception," Int. J. Multilingual. **13**(4), 459–475.
Qin, Z., and Jongman, A. (**2016**). "Does second language experience modulate perception of tones in a third language?," Lang. Speech **59**(3), 318–338.
R Development Core Team (**2018**). "R: A language and environment for statistical computing [computer program]," Vienna, Austria: R Foundation for Statistical Computing. http://www.r-project.org (Last viewed July 12, 2019).
Remijsen, B. (**2016**). "Tone," in *Oxford Research Encyclopedia of Linguistics*, edited by M. Aronoff (Oxford University Press, Oxford, UK).
Rothman, J. (**2011**). "L3 syntactic transfer selectivity and typological determinacy: The typological primacy model," Second Lang. Res. **27**(1), 107–127.
Rothman, J. (**2015**). "Linguistic and cognitive motivations for the Typological Primacy Model (TPM) of third language (L3) transfer: Timing of acquisition and proficiency considered," Bilingual. Lang. Cogn. **18**(2), 179–190.
Rothman, J., and Cabrelli Amaro, J. (**2010**). "What variables condition syntactic transfer? A look at the L3 initial state," Second Lang. Res. **26**(2), 189–218.
Schütze, C. T., and Sprouse, J. (**2014**). "Judgment data," in *Research Methods in Linguistics*, edited by R. J. Podesva and D. Sharma (Cambridge University Press, London, UK), pp. 27–50.
Slabakova, R., and del Pilar García Mayo, M. (**2015**). "The L3 syntax-discourse interface," Bilingual. Lang. Cogn. **18**(2), 208–226.
So, C. K., and Best, C. T. (**2010**). "Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences," Lang. Speech **53**(2), 273–293.
Strange, W. (**2011**). "Automatic selective perception (ASP) of first and second language speech: A working model," J. Phon. **39**(4), 456–466.
Teeranon, P. (**2008**). "The change of Standard Thai high tone: An acoustic study and a perceptual experiment," SKASE J. Theor. Linguist. **5**(1), 1–17.
Thepboriruk, K. (**2009**). "Bangkok Thai tones revisited," Working Papers Linguist.: Univ. Hawai'i Mānoa **40**(5), 1–15.
Tsukada, K., and Kondo, M. (**2019**). "The perception of Mandarin lexical tones by native speakers of Burmese," Lang. Speech (published online).
Tsukada, K., Xu, H. L., and Rattanasone, N. X. (**2015**). "The perception of Mandarin lexical tones by listeners from different linguistic backgrounds," Chin. Second Lang. Res. **4**(2), 141–162.
Wang, X. (**2006**). "Perception of L2 tones: L1 lexical tone experience may not help," in *Speech Prosody 2006: Proceedings of the 3rd International Conference on Speech Prosody*, edited by R. Hoffmann and H. Mixdorff (TUDpress, Dresden, Germany), pp. 85–88.
Wang, X. (**2013**). "Perception of Mandarin tones: The effect of L1 background and training," Modern Lang. J. **97**(1), 144–160.
Wayland, R. P., and Guion, S. G. (**2004**). "Training English and Chinese listeners to perceive Thai tones: A preliminary report," Lang. Learn. **54**(4), 681–712.
Wrembel, M. (**2010**). "L2-accented speech in L3 production," Int. J. Multilingual. **7**(1), 75–90.
Wrembel, M. (**2012**). "Foreign accentedness in third language acquisition: The case of L3 English," in *Third Language Acquisition in Adulthood*, edited by J. Cabrelli Amaro, S. Flynn, and J. Rothman (John Benjamins Publishing, Amsterdam, The Netherlands), pp. 281–309.
Wrembel, M. (**2015**). *In Search of a New Perspective: Cross-Linguistic Influence in the Acquisition of Third Language Phonology* (Adam Mickiewicz University Press, Poznań, Poland).
Yang, C. (**2019**). "The effect of L1 tonal status on the acquisition of L2 Mandarin tones," Int. J. Appl. Linguist. **29**(1), 3–16.
Yip, M. (**2002**). *Tone* (Cambridge University Press, Cambridge, UK).
Zhang, Y., Kuhl, P. K., Imada, T., Kotani, M., and Tohkura, Y. (**2005**). "Effects of language experience: Neural commitment to language-specific auditory patterns," NeuroImage **26**(3), 703–720.