Authors version of accepted manuscript now published in *Journal of Phonetics:*
https://doi.org/10.1016/j.wocn.2022.101197

Native language experience with tones influences both phonetic and lexical processes when acquiring a second tonal language

Eric Pelzl[1*], Jiang Liu[2], Chunhong Qi[3]

[*]*Corresponding author:* pelzlea@gmail.com
[1] The Pennsylvania State University, University Park, Pennsylvania, USA
[2] University of South Carolina, Columbia, South Carolina, USA
[3] Yunnan Normal University, Yunnan, China

Abstract:

Second language acquisition of lexical tones requires that a learner form appropriate tone categories and bind those categories to lexical representations for fluent word recognition. Research has shown that second language (L2) learners with no previous tone language experience can become highly accurate at identification of tones in isolation, but, even at advanced levels, have difficulty using tones to differentiate real words from nonwords. The present research considers the same skills in L2 learners who *do* have previous tone experience. Using largely the same tasks and stimuli previously used with English speakers in Pelzl, Lau, Guo, & DeKeyser (2021a) ("PLGD21"), we examined the tone identification and (tone) word recognition abilities of thirty-three Vietnamese speakers who had achieved advanced L2 proficiency in Mandarin. Results indicate that Vietnamese speakers experience different tone identification difficulties than English speakers, presumably due to interference from their native language tone categories. However, unlike English speakers in previous studies, Vietnamese speakers did not display differences in lexical decision accuracy for vowel and tone nonwords. These results provide evidence of the complexities of cross-linguistic influence, illustrating that the influence of native language tones can be illuminated by considered perception and acquisition at multiple levels.

Keywords: Vietnamese, Mandarin, tones, cross-linguistic influence, speech learning, second language acquisition

1. INTRODUCTION

Although pitch is a universal feature of spoken language, not every language uses it in the same manner. In lexical tone languages, such as Mandarin or Vietnamese, pitch patterns (acoustic differences in F0 height and contour) distinguish one word/morpheme from another. In contrast, non-tonal languages do not have tones, but may have word stress (e.g., English) or pitch accents (e.g., Japanese), along with paralinguistic uses of pitch for intonation or the expression of emotion. These different experiences with linguistic pitch shape how people perceive and use pitch when learning a second language.

The present study investigates whether experience speaking a tonal first language (L1) influences long-term tone and word learning outcomes in a tonal second language (L2). We examined L1 Vietnamese speakers who had achieved advanced L2 proficiency in Mandarin. In order to allow for close comparison with outcomes from non-tonal L1 speakers, we conducted a conceptual replication of Pelzl et al. (2021a)—a study that targeted English L1 speakers who had also achieved advanced L2 proficiency in Mandarin.

This work provides new evidence of cross-linguistic advantages for L2 lexical tone learning that naturally accrue with L1 tone language experience. These advantages are made clear by considering the influence of L1 tone experience both at the level of L2 tone category formation (accuracy and errors in identification of L2 tone categories), and at the lexical level (processing and explicit knowledge of phonological *tone words*).

1.1 Background: Cross-linguistic influences in tone perception

A large portion of the world's extant languages are tonal languages (Maddieson, 2013; Yip, 2002), but not all tonal languages utilize tone in the same way or to the same extent across the lexicon. For the purposes of the present study, 'tonal languages' are specifically those such as Vietnamese, Thai, and Mandarin, with complex tonal systems that are used pervasively across the lexicon.

The effect of tone language experience, or a lack of that experience, has long been a focus of research on cross-linguistic tone perception and the initial stages of second language tone acquisition (for reviews, see Best, 2019; Pelzl, 2019). In two pioneering studies (Gandour & Harshman, 1978; Gandour, 1983), listeners from a variety of tonal and non-tonal L1s were found to rely most heavily on two cues for tone discrimination, namely F0 *height* and the *direction* of F0. Importantly, which dimension was strongest varied not only between tonal and non-tonal speakers, but also among speakers from differing tonal languages. In other words, perception of F0 is not a simple on-off switch (tonal vs. non-tonal).

Nevertheless, L1 tone experience may confer some general perceptual advantages for novel tones. Schaefer and Darcy (2014) found a hierarchy of overall accuracy in discrimination of Thai tones that seemed to directly reflect the phonological function of pitch among four different L1 groups. Other studies have found that even experience in a tonal *L2* can confer positive influences on discrimination of F0 cues in an additional language (Qin & Jongman, 2015; Wiener & Goss, 2019). Chang et al. (2017) suggest that a general advantage might lie in an ability to normalize F0 among different speakers when identifying novel tones.

Still, tonal language experience is not an absolute advantage. The phonetic and phonological influence of previous experience with specific tone categories can also create challenges for identification and discrimination of novel tones. For instance, Hong Kong

Cantonese listeners tend to perceive high-level and high-falling tones as allophones in their L1, and this may result in misidentifications of high-level and high-falling tones in L2 Mandarin (Hao, 2012; So & Best, 2010). Similarly, Francis et al. (2008) found that both English and Mandarin speakers made numerous identification errors after training with Cantonese tones, but that the nature of those errors differed between the non-tonal and tonal listeners. In other words, the similarities and differences between specific L1 and L2 tones will influence how easy/difficult it is to learn the new tone categories.

1.2 Tonal second language acquisition at lower and higher levels of speech learning

Wong and Perrachione (2007) describe L2 learning as happening along a "phonetic-phonological-lexical" continuum, where "more basic auditory abilities (phoneme discrimination) mediate performance on higher level auditory tasks (word learning)" (p. 566). In the present study, we focus on learning challenges related to the two ends of this continuum, the phonetic and lexical levels (Figure 1), which we will describe as *Tone Category Learning* and *Tone Word Learning*.

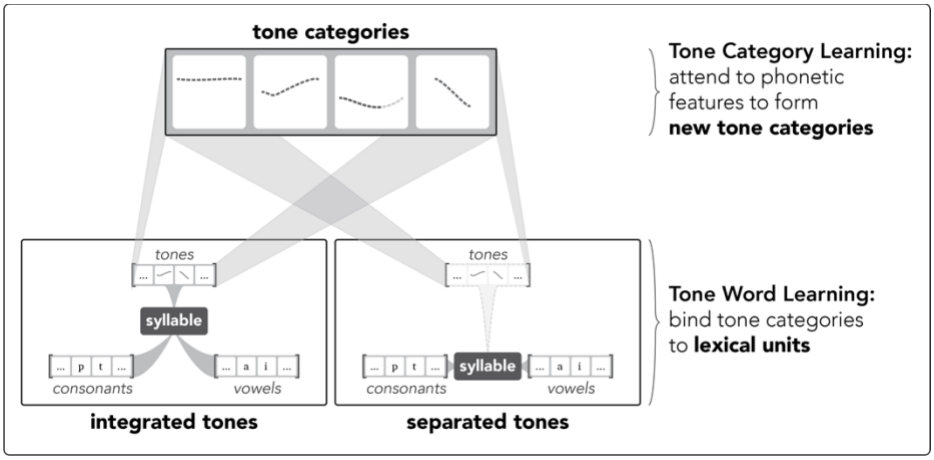

FIGURE 1. Expected challenges for learning a second tonal language. All learners must form novel tone categories for the second language. Whereas native tonal language speakers already have integrated tones, non-tonal language speakers must also learn to integrate tones into lexical units.

When learning new L2 tones, listeners must form appropriate tone categories so that they can smoothly differentiate and identify these tones in spoken input. Insofar as F0 is the primary cue distinguishing the L2's tone categories, the learner's task is to use F0 height and contour differences to form these categories. As noted above, L1 tone experience is not an absolute advantage for these processes, as the interplay of tonal inventories between L1 and L2 may create unique confusions—confusions that a non-tonal L1 speaker may not experience.

Learning tone categories is, however, only one step towards mastery of L2 tones. These categories must also be bound to 'tone words'—the phonological form of lexical representations in long-term memory—so that they can be retrieved in real time for fluent speech comprehension.

While obviously related, these two levels of learning are also different. Any strengths or weaknesses a listener presents in perception of tones in general, or specifically to a tone category, will influence the efficiency of lexical processes. In that sense, both tonal and non-tonal L2 learners face a similar challenge. However, what seems likely to differentiate them is that L1 tone speakers already integrate tones into lexical representations, whereas non-tonal speakers must learn to do so.

1.4 L2 Tone word learning

A number of tone word learning studies have examined how non-tonal language speakers acquire tones in the earliest stages (e.g., Chandrasekaran et al., 2010; Chang & Bowles, 2015; Wong & Perrachione, 2007). Among these studies, that of Cooper and Wang (2012) is most relevant to the present research. Cooper and Wang trained both L1 English and L1 Thai speakers to learn a small set of Cantonese tone words, testing their accuracy both in tone identification

and in matching words to meanings. On the lexical test, the tonal language speakers outperformed the English speakers (though musicians in both groups showed the best performance). The study showed that tonal language speakers had a benefit at the lexical level, even though their performance on phonetic tasks was not clearly superior to that of the non-tonal speakers.

While appropriate to the goals of their study, Cooper and Wang's (2012) focus on naïve listeners and small sets of tonally contrastive words considerably limits what results can tell us about the longer-term outcomes of L2 tone acquisition. Naïve or beginning learners can shed light on the beginning state of L2 tone perception or tone word learning, but cannot show which initial difficulties (or successes) might be superficial, and which more persistent. Perhaps more importantly, the use of minimally contrastive, monosyllabic tone vocabulary (e.g., the syllable /fu/ paired with 5 Cantonese tones) does not reflect of the full lexical complexity that confronts L2 learners. For instance, in the full Mandarin lexicon, there are many 'tone gaps,' syllable-to-tone combinations that never occur (e.g., the syllable /nəŋ/ only ever occurs with T2), thus removing the information value of tones for those words. When words do contrast only in tone, they still rarely overlap in word class (i.e., noun vs. noun). Furthermore, most existing words are disyllabic (Duanmu, 2007), and even monosyllabic words generally do not occur in complete isolation. The multisyllabic nature of words and the information provided by context go a long way to making tones redundant for the listener, and thus undermining their functional importance—especially if the listeners are biased by L1 experience to ignore tones during word recognition.

The functional redundancy of tones may be one reason that recent studies with highly proficient L2 learners from non-tonal language backgrounds consistently find that they have

considerable difficulty using tones lexically (J.-I. Han & Tsukada, 2020; Ling & Grüter, 2022; Pelzl et al., 2019, 2021a, 2021b). Pelzl et al. (2019) found that L1 English speakers who had achieved advanced proficiency in L2 Mandarin typically performed with high accuracy on tone identification tasks, but showed a strong disadvantage for tones compared to vowels in lexical decision tasks. Participants were required to decide whether a disyllabic spoken stimuli was a real word of Mandarin or not. Stimuli consisted of real words and two types of nonword counterparts. In tone nonwords, the spoken item differed from a real word in the tone of its first syllable (e.g., nonword *fa**2**yin1* based on the real word *fa**1**yin1*). For vowel nonwords, they differed in the vowel of the first syllable (e.g. nonword *f**u**1yin1*). Pelzl et al. (2019) used very challenging stimuli that were clipped out of sentences. Average accuracy for vowel nonwords was 84%, but for tone nonwords was 35%, a below-chance score that indicated a bias to accept tone nonwords. Pelzl et al. (2021a) tested a second group of English speakers using the same type of lexical decision task, but with spoken stimuli that had been produced in isolation, resulting in clearer pronunciation and slower speech rate. While overall performance improved—especially for tones—the disadvantage for tone nonwords persisted (vowel nonwords: 85%; tone nonwords: 62%). Similar behavioral results applied for the same L2 participants on a picture-phonology matching task (Pelzl et al., 2021b). Furthermore, vocabulary knowledge test results indicated that participants often misremembered the tones for words, even if they were otherwise correct and confident about the definitions of those same words. Using medium-term repetition priming, J.-I. Han and Tsukada (2020) found similar lexical tone difficulties among L1 Korean speakers with relatively advanced L2 proficiency, who tended to incorrectly accept tone-switched words as repetitions.

These lexical difficulties do not mean that tone category abilities are of no importance for non-tonal L2 learners. Using visual world eye-tracking, Ling and Grüter (2022) also examined experienced L2 learners (L1 English speakers) and found evidence of a disadvantage, relative to native Mandarin speakers, for using tones for real-time tone word recognition. However, they also showed a relationship between the categoricity of L2 learners' perception of tones and their performance on the visual world task. This supports a link between lower level tone category formation and lexical processes. In other words, while these two learning tasks are separable, they are nevertheless closely related. This is important to keep in mind when considering tonal language speakers—tone-specific lower level abilities may have knock-down effects on performance in lexical tasks such that, where the difficult tones are at play, performance may be less robust.

1.4 The relation of Mandarin and Vietnamese tone inventories

Both Mandarin and Vietnamese are tone languages in which lexical tones differentiate word/morpheme meanings. Both languages have level and contour tones. There are four lexical tones in Mandarin (Duanmu, 2007), while Vietnamese[1] has six lexical tones (Nhan, 1984; Yip, 2002). Figure 2 illustrates the tonal categories in Mandarin and Vietnamese.

[1] Northern Vietnamese is the standard variety of Vietnamese spoken in Vietnam that has six tones whereas Southern Vietnamese has five tones as *hỏi* and *ngã* are merged (Han, 1969; Nguyễn & Edmondson, 1997) (Han, 1969; Nguyễn & Edmondson, 1997). Most of the Vietnamese participants in the current study were from regions that use Northern Vietnamese. A few participants from the regions that speak Southern Vietnamese reported that they were taught in and spoke primarily Northern Vietnamese throughout their formal education.

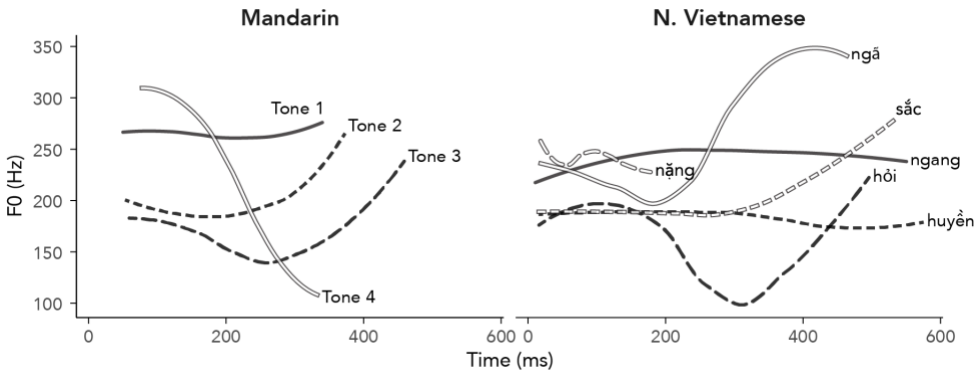

FIGURE 2. Comparison of Mandarin and Vietnamese Tones. Based on recordings of two women, one L1 Mandarin speaker, one L1 Vietnamese speaker. Both produced all the tones of their L1 on the syllable /ma/.

For convenience and consistency with previous research, we label the Mandarin tones with numbers: T1 (high-level), T2 (high-rising), T3 (low, or low-dipping in isolation) and T4 (high-falling). Vietnamese tones are labeled with the traditional Vietnamese phonological terms *ngang* (high-level), *huyền* (low level), *hỏi* (falling-rising), *ngã* (broken falling-rising with glottalization), *sắc* (rising), and *nặng* (short-falling with a glottal stop).

At least three factors may affect L1-Vietnamese L2-Mandarin learners' perception of Mandarin tones. First, based on similarities between Mandarin and Vietnamese tones, three of the Mandarin tones might be relatively easy for Vietnamese participants to categorize. The high-level T1 might be mapped onto one of level tones, *ngang* or *huyền*; the high-rising T2 might be mapped onto the similar rising *sắc*; and the low-dipping T3 might be mapped onto the complex contour tone *hỏi*. However, no Vietnamese tone has a steep falling pitch comparable to that of T4. It is thus likely to be a poor exemplar and L1 Vietnamese learners of Mandarin may have some difficulty forming the relevant T4 category.

A second factor that could explain Vietnamese-speaking learners' perception of Mandarin tones is the cue-weighting differences between Vietnamese and Mandarin. As mentioned above, studies of cross-linguistic tone perception have high-lighted the different perceptual weight that listeners give to F0 cues based on their L1 tone experience (Francis et al.,

2008; Gandour, 1983; Gandour & Harshman, 1978). Native Mandarin speakers primarily rely on pitch contour for categorizing four Mandarin tones (Gandour, 1983). For Vietnamese, with its larger tonal inventory, F0 onset has been found to be weighed equally with pitch height and contour in categorizing the six Vietnamese tones (Brunelle, 2009). As T1 and T4 have very similar F0 onset, and there is no similar falling contour among Vietnamese tones, L1 Vietnamese cue-weighting tendencies may induce difficulty in differentiating T1 and T4.

These expectations of difficulty for T1 and T4 have some empirical support. Tsukada (2019) tested naïve Vietnamese listeners discrimination of Mandarin tone pairs. She found Vietnamese listeners had the greatest difficulties with T1 and T4. Previous research has shown that beginner level L1-Vietnamese L2-Chinese learners numerically had more errors in producing T1 and T4 in L2 Chinese (Wu & Hu, 2004).

A third factor that might impact Vietnamese acquisition of Mandarin tones is L2 proficiency. Typically, we expect that increasing L2 proficiency will coincide with improved L2 perception.

Together, the factors outlined above, as well as previous empirical studies, motivate the following predictions for L1 Vietnamese speakers' ability to identify T1 and T4. First, if Vietnamese participants perceive T1, but not T4, as being a good fit to their L1 tone categories, then we might see an asymmetrical pattern of confusion, such that T1 is identified with high accuracy, but T4 is often identified as T1.[2] Alternatively, if both tones are assimilated to the

[2] One reviewer pointed out that the absence of this confusion would not necessarily be indicative of assimilation of T4 to T1. Vietnamese listeners might detect *phonetic* differences between Vietnamese and Mandarin tones, leading to formation of a new category in their common L1-L2 phonological space. This new category then might preclude misidentification of T4 as T1.

same Vietnamese tone category, or if L1 Vietnamese cue-weighting favors F0 onsets to identify tones, then we should observe similar confusion for both T1 and T4. Finally, if L2 proficiency leads to improvement, we may see few errors among participants, or discernable trends for increased tone identification accuracy as L2 proficiency increases.

In addition to cross-linguistic tone influences, the nature of the tones within the Mandarin inventory must also be kept in mind. Regardless of any listener's previous tonal experience, not all tonal distinctions are equally *distinctive*. In the case of Mandarin, isolated productions of T2 and T3 are often found to be somewhat confusable, even for native listeners (e.g., Huang & Johnson, 2010; Shen & Lin, 1991). Among the Mandarin tones, T3 has the most allophonic variation. In connected speech, instead of being realized as a dipping (falling-rising) F0 contour, T3 is most often realized with a low-falling F0 contour when followed by T1, T2, or T4 (Xu, 1997; Zhang & Lai, 2010). In context then, the contour of T3 often resembles that of T4 (Gårding et al., 1986). T3 also undergoes sandhi when followed by another T3 and is then realized with a rising F0 that appears to be the same as that of T2 (Duanmu, 2007; Zhang & Lai, 2010). These allophones might make T3 more challenging for L1 Vietnamese learners and could result in similar patterns of difficulty for T3 as those observed in previous studies with other L2 groups (Hao, 2012; Pelzl, 2018).

However, another possibility exists for T3. Because Vietnamese has a dense set of dipping and rising tones, Vietnamese participants may display the type of advantageous L1 influence found in some previous cross-linguistic studies (Bohn & Best, 2012; Chang & Mishler, 2012; Wiener & Goss, 2019). That is, due to an increased sensitivity to F0 cues in dipping and rising tones, L1 Vietnamese participants may show little confusion for T3, and perform as well as, or better than, L1 Mandarin participants.

Our predictions were formulated with the assumption that F0 is the primary cues listeners will use to identify Mandarin tones. However, there are secondary cues that might also come into play. Phonation type is another feature of (Northern) Vietnamese tones. In canonical form, *hỏi*, *ngã*, and *nặng* display laryngealization, and this has been found to contribute strongly to L1 Vietnamese tone categorization (Brunelle, 2009). In Mandarin, T3 sometimes also displays creakiness (Kuang, 2017), especially in isolation. Both Vietnamese and Mandarin tones also display predictable durational differences. These secondary cues might aid listeners in identifying tones. We did not formulate predictions based on phonation or duration cues, but we will return to this issue in the discussion.

1.6 The relation of Mandarin and Vietnamese words

No previous studies have examined tone word learning by L1 Vietnamese participants. As described above, due to their experience using F0 as a lexically functional cue, we expect Vietnamese participants to automatically integrate tones into lexical representations, and thus to show similar accuracy for both tone and vowel nonwords in the present study. However, we should note some relationships between Vietnamese and Mandarin that could affect L2 word learning.

Historically, the Vietnamese lexicon was heavily influenced by Chinese. Importantly, the greatest level of influence came via written Chinese, rather than spoken (Alves, 2009). Using estimates based on Vietnamese dictionaries, Alves suggests that as much as 70% of modern Vietnamese vocabulary is Sino-Vietnamese borrowings. Though in most cases the pronunciation of the borrowings is fully Vietnamese, there are detectable correspondences between the languages. Educated Vietnamese speakers will be able to recognize and utilize connections

between Chinese and Vietnamese vocabulary. Given the scope of borrowings, there is no doubt that these loanwords influence Mandarin language learning for Vietnamese speakers—whether for good or ill.

While we acknowledge the potential of such influences, the present study will not attempt to account for them, as our primary goal was to make a direct comparison with Pelzl et al. (2021a).

2. MATERIALS AND METHODS

2.1 Participants

We recruited 58 native Vietnamese-speaking participants who had achieved relatively advanced L2 proficiency in Mandarin. Participants were recruited from two universities in China. At the time of the study, 23 of these participants had studied in China for 6 months or longer; 10 had successfully been admitted to Chinese universities, but had not yet relocated to China. Most (56 out of 58) had passed HSK 5 or 6 (a standardized Mandarin proficiency test administered worldwide from level 1 to 6). In order to provide the best comparison possible with Pelzl, Lau, Guo, and DeKeyser (2021a) (from now on PLGD21), we screened participants using the same proficiency tasks (translated into Vietnamese) and inclusion criteria as in PLGD21. All Vietnamese participants completed a Yes/No Vocabulary assessment and a Can-Do assessment (see section A5 in the supplementary materials for details). Of the 58 Vietnamese participants originally recruited for the study, 44 passed the screening tests. Of these, eleven more were excluded for the following reasons: two for not following instructions; four due to beginning learning Mandarin before age ten; four due to Cantonese heritage language status; one due to missing data (~50% of lexical decision responses missing). This left a final sample of 33

Vietnamese speaking participants who passed the same screening criteria for 'advanced' L2 Mandarin status as used in PLGD21. A summary of Vietnamese participant's background characteristics is shown in Table 1 (for comparison with PLGD21, see supplementary materials A3). Of the 33 participants in the study, 10 had spent no time living or studying Mandarin in China. We will return to the role of study context in the discussion.

A group of 17 L1 Mandarin participants also completed a subset of the experimental tasks to allow for comparison with Vietnamese results.

TABLE 1. Background information and screening measures for Vietnamese participants (*n*=33)

	mean (sd)	range
Age at testing	25 (3.4)	20-34
Age of onset	19.3 (2.3)	16-26
Years in immersion	1.6 (2.0)	0-7
Total years learning	5.7(3.4)	2-16
Can-do self-assessment (%)	82.1 (7.2)	72-100
Vocabulary self-assessment (%)	94.5 (4.1)	85-100
HSK score [a]	5.5 (0.51)	5-6

[a] *Two participants had not taken the HSK and so are not included in this estimate*

2.2 Stimuli

2.2.1 Tone identification stimuli

One key goal of this identification task was to discourage ceiling performance. To this end, steps were taken to ensure the difficulty of the task by (1) using nonword stimuli, (2) multiple talkers, and (3) multiple target contexts: monosyllables, disyllables, and clipped syllables.

Four pronounceable non-Mandarin syllables were selected (*bou* /pəu/; *chei* /ʈʂʰəi/; *fai* /fai/; *tiu* /tʰiəu/). Each syllable was combined with all four tones to create a set of 16 monosyllables. Sixteen disyllables were then created by adding the syllable /pa/ following the

first syllable. This second syllable was always unstressed and produced with a neutral tone (*qing sheng*). The neutral tone was chosen for the second syllable to minimize coarticulatory influences on the tone of the first syllable (Chen & Xu, 2006; Lee & Zee, 2008). The syllable /pa/ was selected because its tone is often neutralized in disyllabic words in standard Mandarin (e.g., *zui3ba* 'mouth').

Ten native Mandarin speakers (3 men, 7 women) recorded all 32 targets on the campus of Beijing Normal University. Speech was recorded in 16 bits at a 44.1 kHz sampling rate using Praat (Boersma & Weenink, 2018). While standing in a sound-attenuated booth, each speaker read the syllables presented in Pinyin on cue cards that were presented in a random order for each speaker. After all recordings were completed, four speaker (2 men, 2 women) were selected to use for the experiment based on listener perceptions of having relatively mild or no laryngealization in their productions of T3 in isolation. This selection process was meant to minimize the use of creakiness as a secondary cue for identifying T3. Next, clipped syllable stimuli were created in Praat by extracting the first syllable of each disyllable item prior to the closure of the stop in the /pa/ syllable. This resulted in an additional 16 clipped syllable stimuli per speaker. In total, there were 192 unique auditory stimulus tokens (4 speakers × 4 syllables × 4 tones × 3 contexts). Stimulus intensity was normalized to 70 dB, no other normalization was applied. The stimuli are the same as those used with participants from PLGD21, except that the clipped syllables were not included when L2 English speakers were tested.

The inclusion of MS, DS, and CS syllables was motivated by a desire to test the effects of context on the identification of Mandarin tones. Due to their sometimes similar falling-rising contours, T2 and T3 are often confused in isolation (Shen & Lin, 1991). However, when T3 occurs in context, it typically loses the dipping contour and is better described as a low-falling

tone (Gårding et al., 1986). This low-falling contour is more similar to T4. An illustrative example of F0 contrasts between MS and DS is shown in Figure 3. Given these allophonic realizations of Tone 3, we reasoned that we would see qualitatively different response patterns for each context. In MS context, T3 would be confused with T2. In DS context, T3 would be identified with high accuracy given that it no longer resembles T2 and this is the most natural form in which T3 occurs. In CS context, we reasoned that by removing the contextualizing F0 of the following syllable, T3 might be confused with T4. Given that these expectations were all phonetically motivated, we expected that both Vietnamese and Mandarin listeners would display similar patterns of accuracy and error for T3 in each context.

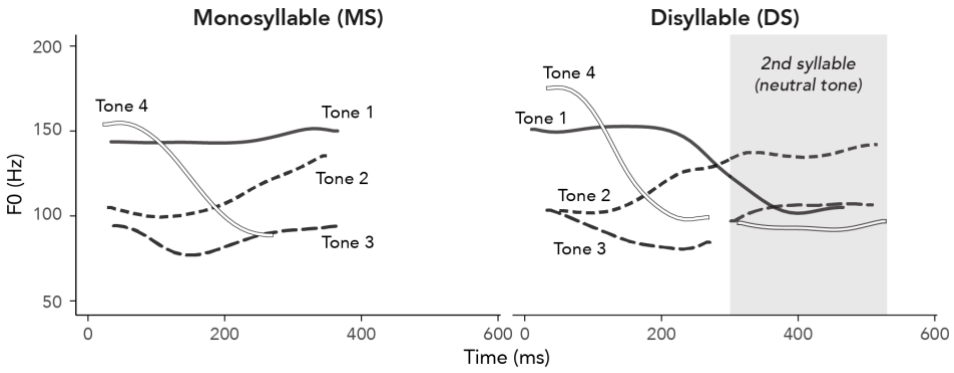

FIGURE 3. Comparison of Mandarin monosyllable and disyllable stimuli. Based on recordings from one man, an L1 Mandarin speaker, for the syllables /pou/ and /pou-pa/. Clipped syllables were created by removal of the 2nd syllable of disyllable stimuli.

2.2.2 Lexical decision stimuli

Lexical decision stimuli were the same as those used in PLGD21. Ninety-six high frequency real words were selected. For each real word, a vowel and tone nonword were created by changing the vowel or tone on the first syllable. For example, based on the real word *bai2tian1* /pai2 tʰien1/ (“daytime”), we created vowel nonword *ba2tian1* /p**a**2tʰien1/, and tone nonword *bai**3**tian1* /pai3 tʰien1/. The occurrence of tones in real words, and their replacement in nonwords

was balanced to represent all four tones as evenly as possible. Vowels were replaced with a wide variety of other vowels and care was taken so that vowel replacements in nonwords respected the phonotactics of Mandarin. Three counter-balanced lists were created so that each participant would be tested on 32 real words, 32 vowel nonwords, and 32 tone nonwords. Additionally, 32 filler real words were included to balance the proportion of yes/no responses across the task. For additional details, see Pelzl et al. 2021a. The list of all stimuli, including audio files, is available on the Open Science Forum (DOI 10.17605/OSF.IO/VE6PZ).

2.3 Procedures

The procedures was modeled closely on those of PLGD21, except that no electroencephalogram (EEG) was recorded, and all tasks were administered via web-based experimental platforms Ibex (Alex Drummond) and PCIbex (Zehr & Schwartz, 2018). Each Vietnamese participant completed seven tasks (Figure 4). First, they provided brief background language history, then they completed the Can-Do assessment and Vocabulary self-assessment tests. Immediately following the self-assessment tests, they proceeded to the lexical decision task, and then continued on to the tone identification task. After tone identification, participants also completed a tone word knowledge test that checked their knowledge of the vocabulary used in lexical decision, and finally a brief tone knowledge survey that explored their understanding of the Mandarin tone categories and their opinions about which tones were easy/difficult to learn.

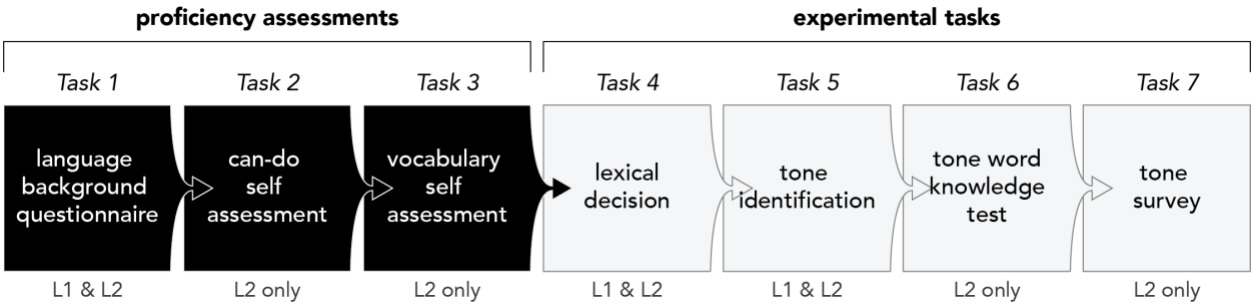

FIGURE 4. Order of tasks. L1 Mandarin participants completed only Tasks 1, 4, and 5.

All instructions were presented in written Vietnamese (or Chinese for L1 Mandarin participants). Before starting the experimental tasks, participants were reminded to wear headphones and checked a box to indicate they were doing so. They then completed a volume check with a non-linguistic pure tone stimulus.

In order to avoid drawing special attention to tones prior to the lexical decision task, lexical decision was administered before the tone identification task. However, for ease of presentation and discussion, in the rest of this article, we present tone identification first followed by lexical decision and other tasks.

2.3.1 Tone Identification Procedures

Participants were told they would hear nonword syllables with the four Mandarin tones and that they should identify the tone of each syllable, or, in the case of disyllables, only the first syllable. Pictures illustrated the monosyllabic and disyllabic stimuli and indicated that only the first syllable of the disyllabic stimulus should be identified. Participants were shown an image illustrating how to place their fingers on the number keys for responses (for images, see supplemental materials A1 and A2).

The three conditions (monosyllables, disyllables, clipped syllables) were presented in counter-balanced order across participants. Each block began with a brief reminder to identify

the tones, followed by four practice trials without feedback. For the disyllabic block, participants were reminded to only identify the first syllable and, in order to proceed, had to click a checkbox acknowledging that they understood they were only to identify the tone on the first syllable. During the task they saw the four tone numbers displayed along with a progress bar. There was a one second inter-trial interval between the end of the previous trial and the beginning of the auditory event in the next trial. Participants had 5 seconds to respond to each auditory stimulus. If they responded, it would cue the inter-trial interval for the next trial, otherwise this would happen automatically after 5 seconds had elapsed.

Each participant completed three blocks of 64 trials, that is, a total of 192 trials. At the end of the task, their results were sent to the server and they clicked on a link to proceed to the next part of the study.

2.3.2 Lexical Decision Task Procedures

Instructions explained that participants would hear real and fake Mandarin words. For each item they were instructed to decide whether it was or was not a real Mandarin word. They were told to place their fingers on the F and J keys as illustrated in a photo.

Throughout the task, the screen displayed reminders that F was ‘yes’ (是) and J was ‘no’ (否), and a progress bar was displayed at the top of the screen. There was a one second inter-trial interval between the end of the previous trial and the beginning of the auditory event in the next trial. Participants had 5 seconds to respond to each auditory stimulus. If they responded, it would cue the inter-trial interval for the next trial, otherwise this would happen automatically after 5 seconds had elapsed.

Participants first completed 14 practice trials without feedback, then completed 128 experimental trials in two blocks of 64, with a self-paced break in between. At the end of the task results were submitted to the server and participants were shown a link to continue to the next part of the study.

2.3.3 Tone word knowledge test

After completing the auditory tasks, participants completed an untimed tone word knowledge test. In this test, they provided tones and definitions as well as confidence ratings for the tones and definitions of all 96 critical real words in the lexical decision task. This test was meant to ascertain whether participants knew the real words that served as the basis of the vowel and tone nonwords; whether they knew the correct tones for those words; and how much confidence they had in their knowledge of the definitions and tones of the words. For each item they supplied the tones with two numbers (e.g., 1 4 would mean T1 and T4), a tone confidence score from 0-3, a definition in Vietnamese, and a definition confidence score from 0-3. The meaning of the confidence scores were shown continuously (in Vietnamese) on the screen:

0 = I don't recognize this word

1 = I recognize this word, but am very uncertain of the tones/meaning

2 = I recognize this word, but am a bit uncertain of the tones/meaning

3 = I recognize this word, and am certain of the tones/meaning

2.3.4 Tone survey

The final task in the study was a short set of questions assessing participant's explicit knowledge about Mandarin tones. Questions and answers were provided in Vietnamese (the list of questions can be found in the supplementary materials A3).

3. RESULTS

3.1 Results of tone identification

Descriptive results of mean accuracy indicate substantial variability among both abroad and home Vietnamese participants (Table 2; Figure 5). Before considering inferential statistical analyses, we can glean several insights from careful consideration of the raw data, especially by looking at individual Vietnamese participant outcomes, which are depicted as shaded dots in Figure 5.

TABLE 2. Accuracy results for Tone Identification

Group	Context	Tone	Mean %	(SD)
Vietnamese (*n*=33)	MS	T1	80.6	(39.6)
		T2	84.4	(36.3)
		T3	77.2	(42.0)
		T4	83.5	(37.1)
	DS	T1	69.1	(46.2)
		T2	87.2	(33.4)
		T3	88.4	(32.0)
		T4	75.0	(43.3)
	CS	T1	66.2	(47.4)
		T2	72.5	(44.7)
		T3	43.5	(49.6)
		T4	71.7	(45.1)
Mandarin (*n*=17)	MS	T1	98.2	(13.5)
		T2	95.6	(20.6)
		T3	80.8	(39.5)
		T4	97.8	(14.7)
	DS	T1	93.0	(25.5)
		T2	87.4	(33.2)
		T3	94.9	(22.1)
		T4	96.7	(18.0)
	CS	T1	93.7	(24.3)
		T2	91.1	(28.5)
		T3	46.5	(50.0)
		T4	96.3	(19.0)

MS = monosyllable; DS = disyllable; CS = clipped syllable

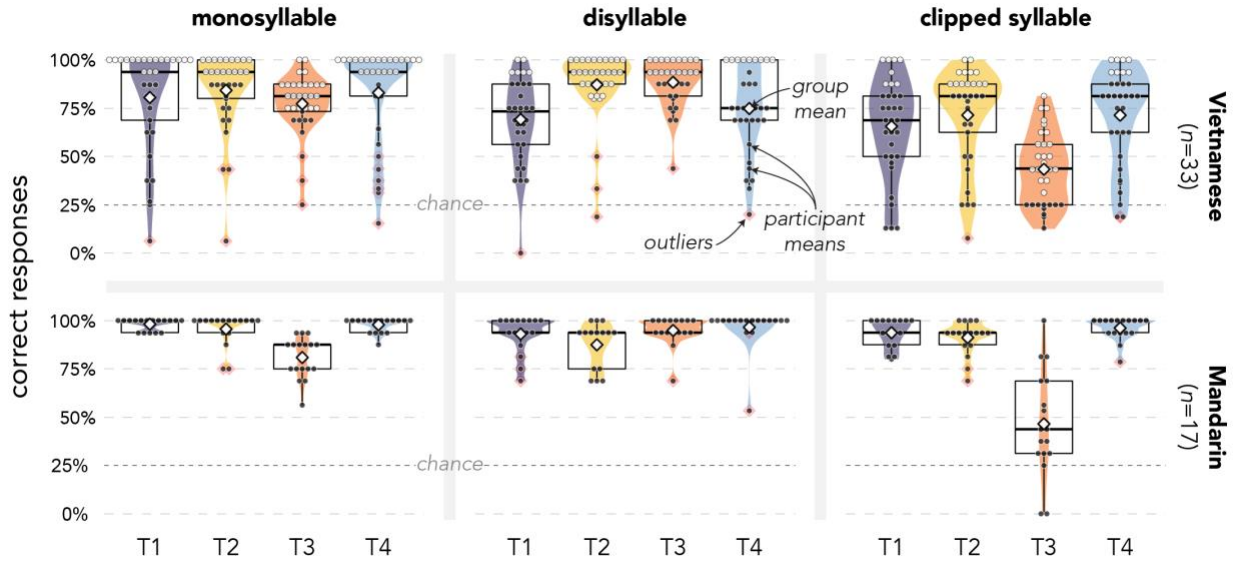

FIGURE 5. Raw accuracy for tone identification. Group means are depicted by white diamonds. Individual participant means are depicted by circles, a red diamond behind the circle indicates an outlier. Vietnamese participants who scored with "native-like" accuracy, i.e., within or above the interquartile range of Mandarin participant scores, are depicted as white circles.

For monosyllabic (MS) stimuli, mean accuracy of both Vietnamese groups was roughly 10% or more below the L1 Mandarin scores for three out of four Mandarin tones (T1, T2, and T4). Nevertheless, when considering individual performance, it is apparent that many or even most Vietnamese participants performed quite well for all monosyllabic stimuli. This can be seen in Figure 5, where Vietnamese participants who scored within the interquartile range of L1 Mandarin participants are depicted as white circles. Over half of all Vietnamese participants performed at native-like levels for all tones in the monosyllable context. While performance dipped for T3, T3 performance also dipped for L1 Mandarin participants.

When considering disyllabic (DS) stimuli, descriptive results for T1 and T4 depict the Vietnamese group as less accurate compared to the Mandarin group. Compared to MS stimuli, fewer individual Vietnamese performed in the range of L1 Mandarin scores for T1 and T4. Only six Vietnamese were in the L1 range for T1 in disyllables, and only nine for T4 in disyllables.

This contrasts with T2 and T3, where the majority of Vietnamese participants were within the L1 range for accuracy.

For clipped syllables (CS), the most striking trend is that most participants frequently misidentified T3 stimuli. This was true regardless of group. In fact, two Mandarin participants misidentified *all* T3 stimuli. Compared to the Mandarin group, the Vietnamese group was less accurate on T1, T2, and T4 in CS.

One final observation is that a number of Vietnamese participants performed very poorly on some or all tones. Such poor performance could indicate tone identification difficulties, but might also indicate other issues. For example, one Vietnamese participant performed below chance on most tones in all conditions. From examination of this participant's post-experiment survey responses, it appears they did not understand the standard Mandarin tone labels correctly (i.e., their description of each Mandarin tone's contour describes one of the other tones, rather than the appropriate tone).[3]

Accuracy results for tone identification were fit to mixed-effects logistic regression models using *lme4* (version 1.1.21, Bates, Mächler, Bolker, & Walker, 2015; using the *bobyqa* optimizer) in *R* (version 4.0.3, R Core Team, 2020). The dependent variable was Accuracy (1,0). Fixed effects included the factors Group (Mandarin, Vietnamese), Tone (T1, T2, T3, T4),

[3] On the advice of a reviewer, we conducted two additional analyses of the tone identification. First, we excluded the participant who confused tone labels; second, we excluded the five lowest performing Vietnamese participants. These exclusions both resulted in a failure to find significance for the difference between Mandarin and Vietnamese groups in accuracy of T4 in MS. No other results were substantively different. These changes suggest that the outcomes for T4 in the MS context were strongly affected by these low-performing individuals, but we have no *a priori* reason to exclude them from consideration as representative learners from this population.

Context (monosyllable, disyllable, clipped syllable), and their interactions. Random effects included subjects, syllables, and talkers. Effects coding was applied using the *mixed* function in *afex* (version 0.28-0, Singmann et al., 2020). The maximal random effects model was fit first (Barr et al., 2013; Bates et al., 2015). The best-fitting model with no singular fit warnings was determined by model comparison conducted through likelihood ratio tests, building from the maximal model (which was rejected due to convergence issues) to progressively less complex models. The final model included by-subject random intercepts and slopes for the effect of tone, and by-item random intercepts for items (*glmer model formula*: accuracy ~ context * tone * group + (1 + context + tone | subject) + (1 | syllable) + (1 | talker).

Table 3 reports a mixed model ANOVA table for the tone identification. P-values were obtained using the likelihood ratio test ("LRT") method (additional model details are reported in the supplementary materials).

TABLE 3. Mixed Model ANOVA Table for Tone ID accuracy results (Type 3 tests, LRT-method)

Effect	Df	Chisq.	Chi Df	Pr(>Chisq)	
Context	45	48.68	2	<.001	***
Tone	44	42.95	3	<.001	***
Group	46	23.01	1	<.001	***
Context × Tone	41	185.54	6	<.001	***
Context × Group	45	0.36	2	.835	
Tone × Group	44	23.74	3	<.001	***
Context × Tone × Group	41	26.01	6	<.001	***

Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1

Table 4 presents post-hoc comparisons of interest. Comparisons were specified using the *multcomp* (Hothorn et al., 2008) and *emmeans* (Lenth, 2022) packages. In the MS, DS, and CS contexts, the Vietnamese group was significantly less accurate than the Mandarin group on T1 and T4. Additionally, in the CS context, the Vietnamese group was significantly less accurate than the Mandarin group on T2.

Compared to its own performance in the MS context, the Vietnamese group was less accurate for T1 and T4 in the DS context, and more accurate for T3 in the DS context. They were also less accurate for all tone categories in CS context than in DS context. Finally, the Vietnamese group was less accurate on both T2 and T3 in CS compared to DS contexts. In contrast, for the Mandarin group, only accuracy for T3 varied across contexts. The Mandarin group was more accurate in DS than MS context. Compared to both MS and DS contexts, they were much less accurate in the CS context.

TABLE 4. Post-hoc comparisons for accuracy results in the Tone ID

						95% CI	
Comparison	Estimate	SE	z	Pr(>|z|)		*lower*	*lower*
Between Group Comparisons							
Monosyllables							
Vietnamese vs Mandarin MS T1	2.51	0.60	4.18	.001	***	0.62	4.39
Vietnamese vs Mandarin MS T2	1.31	0.49	2.67	.121		-0.24	2.86
Vietnamese vs Mandarin MS T3	0.19	0.31	0.61	1.000		-0.80	1.18
Vietnamese vs Mandarin MS T4	2.10	0.68	3.10	.035	*	-0.04	4.24
Disyllables							
Vietnamese vs Mandarin DS T1	2.05	0.43	4.80	<.001	***	0.70	3.39
Vietnamese vs Mandarin DS T2	-0.06	0.43	-0.15	1.000		-1.41	1.28
Vietnamese vs Mandarin DS T3	0.88	0.40	2.20	.366		-0.38	2.14
Vietnamese vs Mandarin DS T4	2.79	0.58	4.83	<.001	***	0.96	4.61
Clipped Syllables							
Vietnamese vs Mandarin CS T1	2.28	0.45	5.08	<.001	***	0.86	3.70
Vietnamese vs Mandarin CS T2	1.50	0.42	3.58	.007	**	0.18	2.83
Vietnamese vs Mandarin CS T3	0.15	0.26	0.59	1.000		-0.66	0.97
Vietnamese vs Mandarin CS T4	2.52	0.58	4.32	<.001	***	0.68	4.37
Between Contexts Comparisons							
Monosyllables vs Disyllables							
Vietnamese T1 MS vs DS	0.94	0.22	4.23	.001	***	0.24	1.64
Vietnamese T2 MS vs DS	-0.22	0.25	-0.87	1.000		-1.01	0.58
Vietnamese T3 MS vs DS	-0.90	0.23	-3.89	.002	**	-1.64	-0.17
Vietnamese T4 MS vs DS	1.03	0.24	4.26	<.001	***	0.27	1.78
Mandarin T1 MS vs DS	1.40	0.54	2.60	.142		-0.30	3.10
Mandarin T2 MS vs DS	1.16	0.41	2.82	.082	.	-0.14	2.45
Mandarin T3 MS vs DS	-1.59	0.38	-4.20	.001	***	-2.79	-0.40
Mandarin T4 MS vs DS	0.34	0.59	0.58	1.000		-1.51	2.19
Monosyllables vs Clipped Syllables							

Vietnamese T1 MS vs CS	1.09	0.20	5.41	<.001	***	0.45	1.72
Vietnamese T2 MS vs CS	1.01	0.21	4.76	<.001	***	0.34	1.67
Vietnamese T3 MS vs CS	1.73	0.18	9.60	<.001	***	1.16	2.30
Vietnamese T4 MS vs CS	1.11	0.22	5.01	<.001	***	0.41	1.80
Mandarin T1 MS vs CS	1.31	0.52	2.50	.175		-0.35	2.96
Mandarin T2 MS vs CS	0.81	0.40	2.05	.487		-0.44	2.07
Mandarin T3 MS vs CS	1.77	0.25	7.02	<.001	***	0.97	2.57
Mandarin T4 MS vs CS	0.69	0.54	1.26	1.000		-1.03	2.40
Disyllables vs Clipped Syllables							
Vietnamese T1 DS vs CS	0.15	0.21	0.70	1.000		-0.52	0.82
Vietnamese T2 DS vs CS	1.22	0.24	5.05	<.001	***	0.46	1.99
Vietnamese T3 DS vs CS	2.64	0.23	11.38	<.001	***	1.91	3.37
Vietnamese T4 DS vs CS	0.08	0.23	0.35	1.000		-0.63	0.79
Mandarin T1 DS vs CS	-0.09	0.42	-0.22	1.000		-1.41	1.23
Mandarin T2 DS vs CS	-0.34	0.37	-0.93	1.000		-1.50	0.82
Mandarin T3 DS vs CS	3.36	0.38	8.95	<.001	***	2.18	4.55
Mandarin T4 DS vs CS	0.35	0.54	0.64	1.000		-1.36	2.05

*Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1*
p-values adjusted by Bonferroni-Holm method; asymptotic confidence intervals reported

Examination of error patterns (Figure 6) provides some qualitative context for interpreting tone identification accuracy results. For Vietnamese participants, in all contexts, T1 was most often misidentified as T4 (63% of MS errors; 85% of DS errors; 65% of CS errors), and T4 was most often misidentified as of T1 (64% of MS errors; 80% of DS errors; 62% of CS errors). These error patterns provide evidence that both high-level and falling tones are confusable for many Vietnamese learners of Mandarin.

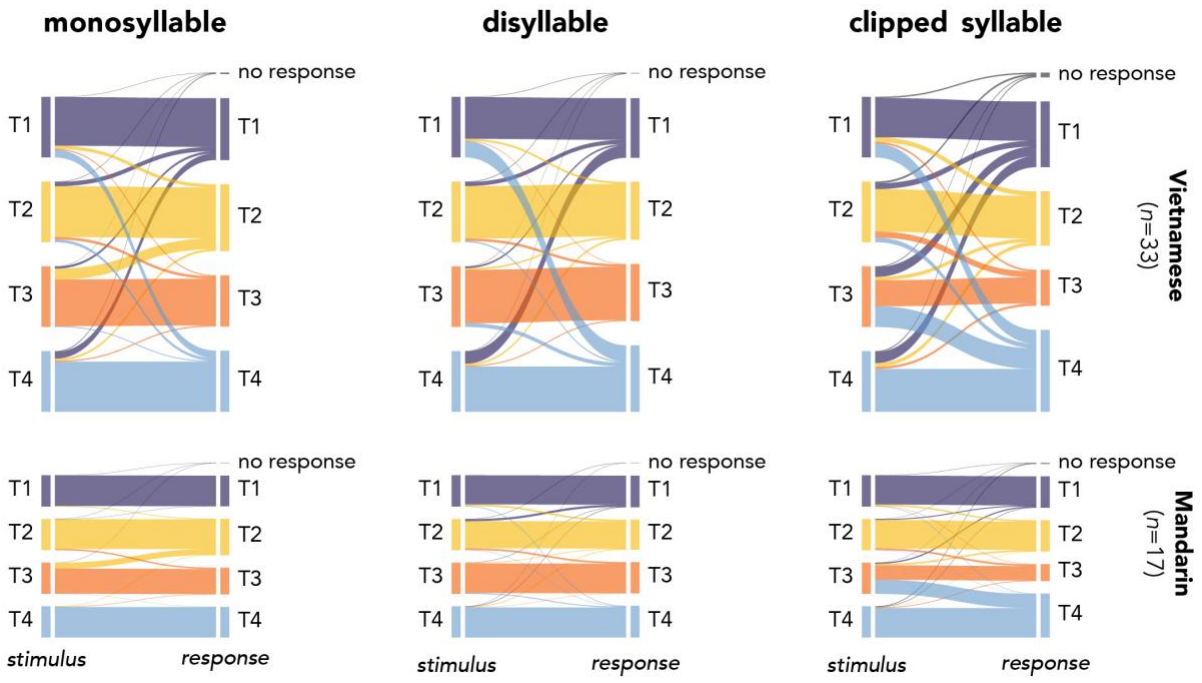

FIGURE 6. Alluvial plots depicting error patterns in the tone identification task. Stimulus tone indicated on left, group responses indicated on right. Thickness of bands indicates total number of responses. Vietnamese plots are thicker overall due to the greater number of participants in that group.

For the low (dipping) T3, Vietnamese learners displayed a similar pattern as native Mandarin listeners, misidentifying the low-dipping T3 as the rising T2 in MS (Vietnamese: 78% of MS T3 errors; Mandarin: 96%), then displaying higher accuracy for T3 in DS. In CS, both groups showed a strong tendency to misidentify the low-falling allophone of T3 as T4 (Vietnamese: 60% of CS T3 errors; Mandarin: 84%); the Vietnamese group additionally showed a tendency to misidentify T3 as T1 (28% of CS T3 errors). Overall, the errors related to T3 supported our predictions, namely, that T3 in MS would be confusable with T2, whereas it would be confusable with T4 in CS. The tendency of the Vietnamese group to misidentify the low-falling T3 in CS as T1 (instead of T4) is consistent with a persistent confusion of T4 and T1.

Finally, though not a pattern we predicted, errors for T2 were also more common for the Vietnamese group in CS, with a relatively even spread of misidentifications across tones (as T1:

31%; as T3: 33%; as T4: 25%; *no response:* 10%). Despite these increased errors, overall, the Vietnamese group was still quite accurate for T2.

3.2 Results of lexical decision

Descriptive results are shown in Table 5 and depicted visually in Figure 7. Overall, native Mandarin listeners were more accurate than Vietnamese listeners, with clear drops in accuracy for Vietnamese responses to nonwords. Of greatest interest in the current study, the Vietnamese group had only a 4% difference in accuracy between vowel nonwords and tone nonwords.

TABLE 5. Descriptive accuracy results for lexical decision task

Group	condition	mean acc. % (sd)
Mandarin (*n*=17)	real	97.6 (15.3)
	vowel	92.9 (25.7)
	tone	93.7 (24.4)
Vietnamese (*n*=33)	real	92.3 (26.5)
	vowel	74.1 (43.8)
	tone	70.1 (45.8)

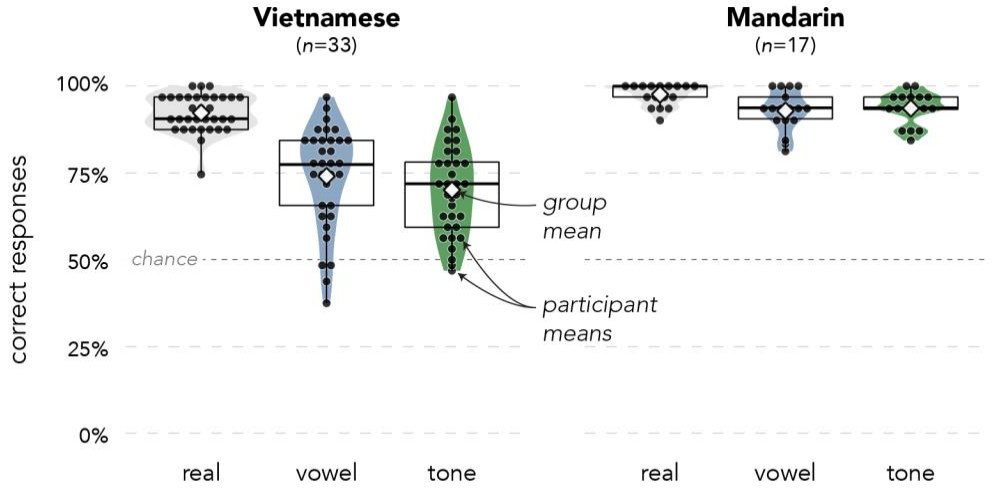

FIGURE 7: Raw accuracy for lexical decision task. Black circles indicate individual participant's mean score in each condition. White diamonds indicate group mean score in each condition.

We also computed d′ as a measure of response bias: $d' = z(\text{H}) - z(\text{F})$ (Macmillan & Creelman, 2005). To correct for hit rates or false alarm rates of 0 or 1, we applied Laplace smoothing (Jurafsky & Martin, 2009). We compared responses for vowel nonwords compared to real words, and tone nonwords compared to real words. Mandarin participants had similar d′ for both conditions (vowel: $m = 3.49$, $sd = .54$; tone: $m = 3.53$, $sd = .49$). Vietnamese participants had overall lower d′ than native speakers, and their group vowel d′ ($m = 2.23$, $sd = .60$) was slightly higher than tone d′ ($m = 2.09$, $sd = .55$). Among all Vietnamese participants, 11 had higher tone d′ than vowel d′, 19 had higher vowel d′, and 3 had equivalent scores (see supplementary materials section A2 for further details and visualization).

Accuracy results were submitted to a mixed-effect logistic regression (using the *bobyqa* optimizer) with crossed random effects for subjects and items. The dependent variable was accuracy (1, 0). Fixed effects included the factors *condition* (real word, tone mismatch, vowel mismatch), and *group* (Mandarin, Vietnamese), and their interaction. The maximal random effects model was fit first (Barr et al., 2013; Bates, Kliegl, et al., 2015). The best fitting model with no singular fit warnings was determined by model comparison conducted through likelihood ratio tests, building from the maximal model (which was rejected due to convergence issues) to progressively less complex models. The final model included by-subject random intercepts and slopes, and by-item random intercepts and slopes for condition and group and their interaction (*glmer* model formula*:* accuracy ~ condition * group + (1 | subject) + (1 + condition + group | item)).

Table 6 reports mixed model ANOVA results for the lexical decision task. There were significant effects of Condition and Group, but the Condition-by-Group interaction failed to reach statistical significance.

TABLE 6. Mixed Model ANOVA Table for LDT accuracy results (Type 3 tests, LRT-method)

Effect	Df	Chisq.	Chi Df	Pr(>Chisq)	
Condition	15	48.74	2	<.001	***
Group	16	41.68	1	<.001	***
Condition × Group	15	3.10	2	.213	

Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1

Table 7 reports post hoc comparisons. Between groups, Vietnamese participants were significantly less accurate than Mandarin participants across all conditions. Within conditions, there was no statistically significant difference between Vietnamese or Mandarin listeners accuracy in rejection of vowel and tone nonwords, while responses to real words were always more accurate than to nonwords (both vowel and tone). Relatively wide CIs indicate some uncertainty around all comparisons.

TABLE 7. Post hoc comparisons for accuracy results in the lexical decision task (p-values adjusted by Holm method)

						95% CI	
Comparison	Estimate	SE	z	Pr(>|z|)		*lower*	*lower*
Vietnamese: Real vs. Vowel	2.13	0.31	6.86	<.001	***	1.28	2.97
Vietnamese: Real vs. Tone	2.24	0.33	6.74	<.001	***	1.34	3.15
Vietnamese: Vowel vs. Tone	0.12	0.22	0.54	.637		-0.47	0.71
Mandarin: Real vs. Vowel	1.89	0.51	3.70	.001	**	0.50	3.27
Mandarin: Real vs. Tone	1.50	0.54	2.79	.016	*	0.04	2.97
Mandarin: Vowel vs. Tone	-0.38	0.38	-1.00	.637		-1.42	0.66
Real: Mand. vs. Viet.	1.70	0.48	3.55	.002	**	0.40	2.99
Vowel: Mand. vs. Viet.	2.44	0.36	6.74	<.001	***	1.45	3.42
Tone: Mand. vs. Viet.	1.94	0.34	5.63	<.001	***	1.00	2.87

*Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1*
p-values adjusted by Bonferroni-Holm method; asymptotic confidence intervals reported

3.3 Results of the vocabulary knowledge test

Vocabulary knowledge test results constitute a rich, but complicated set of data. Here we highlight those results most relevant to our predictions (additional information can be found in the supplementary materials section A2).

3.3.1 Accuracy of definitions

Definitions were scored by two raters (L1 Vietnamese speakers) using a list of anticipated Vietnamese translations of the target vocabulary. Raters first scored all responses according to the list, and also made independent judgments about any alternative translations that were provided. Afterwards, the two raters met and reached agreement on any items where their initial assessment had diverged.

Overall definition accuracy was high (*m* = 93.8%; *sd* = 24.1%; *min* = 56.2%; *max* = 100%). The minimum score is notably lower than what was typical. The next lowest score was 81%. In other words, most participants knew most of the vocabulary included in the lexical decision task.

3.3.2 Accuracy of tones

Tones were scored using the same procedures as in PLGD21. The list of 'correct' tones was based on a predetermined set of tones. For words containing third tone sandhi (T3 followed by T3), both the sequences T2T3 and T3T3 were scored as correct. Each word was given a single score as correct (1) or incorrect (0), that is, no partial credit was given if the tone of one syllable was correct and the other incorrect.

Overall Vietnamese participants provided correct tones for 79.6% of words (*sd* = 40.3%; *min* = 15.8%; *max* = 99%). Even when they were confident of their tone knowledge, they were still incorrect about 15% of the time. Three participants scored below 50% accuracy in providing tones for words. As in the tone identification results, the lowest scoring participant's difficulty seems to have been due to confusion over the appropriate tone labels.

3.3.3 The "Best Case Scenario" for lexical decision

As in PLGD21, we can use the vocabulary knowledge test results to refine the analysis of lexical decision results. By taking the subset of each participants vocabulary test that indicated correct and confident knowledge (ratings of 3) of both the word's tones and its definition, we can see whether this confident and correct knowledge would impact the accuracy of correct rejection for tone and vowel nonwords on the lexical decision task. The intuition is that, because participants knew the tones of these words correctly and confidently, this re-analysis should more strongly affect the results for tone nonwords (whereas confident knowledge of the vowel nonwords would not be impacted by correct/confident knowledge of tones).

Results indicate a small improvement (3-4% increase) in results for both tone and vowel nonwords (vowel nonwords: m = 77.4, sd=41.9; tone nonwords: m=74.6, sd=43.6). Results were submitted to the same statistical modeling procedures as reported above, but without a fixed effect of group (no L1 Mandarin data was included). The final model included by-subject random intercepts and slopes, and by-item random intercepts and slopes for condition (*glmer* model formula): accuracy ~ condition + (1 | subject) + (1 + condition | item). Full model details are included in the supplemental materials.

Model results were consistent with those of the original lexical decision model. The model failed to find a significant difference in accuracy between vowel and tone nonword conditions (b = 0.07; SE = 0.27; z = 0.25; p = .804; *95% CI:* [-0.45, 0.59]).

3.4 Results of the tone survey

At the end of the experiment, all Vietnamese learners of Chinese answered to a set of questions regarding their knowledge about the four lexical tones in L2 Chinese (all questions

available in Appendix A3). First, we asked the learners to describe the pitch shape of the four tones ("*In just a few words, please describe the normal pitch shape of each of the four Mandarin tones.*"). The learners provided descriptions such as "T1 is a high flat pitch", "T2 is from low to high", "T3 is falling first then rising", "T4 is from high to low". Even though we did not ask them to compare the Chinese tones to Vietnamese tones in the questions, most (20 out of 33) provided additional comparisons such as "[T1] sounds like *ngang* in Vietnamese", "[T2] sounds similar to *sắc*", "[T3] is similar to *hỏi*". While such descriptions were common for T1, T2, and T3, for T4, only two learners provided this type of cross-linguistic comparison. One example was, "T4 is sort of a combination of *huyền* and *nặng* in Vietnamese, but not exactly the same." These qualitative responses suggest that Vietnamese listeners tended to map T1, T2, and T3 to their native tone categories, but not T4.

When asked which tone pair they felt was the most confusable in Mandarin, all 33 participants unanimously reported T1 and T4 to be the most confusable tone pair. When asked which single tone was the most difficult, nine learners said that T1 was the most difficult tone to recognize and 21 reported T4 the most difficult to identify. Two learners reported T2 and T3 the most difficult. Altogether, the responses to the tone survey questions indicated that the Vietnamese participants were generally aware of the difficulty identifying T1 and T4. These qualitative responses converge with tone identification results, indicating that T1 and T4 were mutually confusable.

4. DISCUSSION

By considering L2 tone acquisition outcomes at both the level of phonetic category formation and (tone) word recognition, we can see more clearly the ways that L1 experience

influences L2 tone acquisition—both positively and negatively. In the present study, Vietnamese speakers who had achieved advanced proficiency in L2 Mandarin showed evidence of both advantages and disadvantages at the level of phonetic category formation, whereas they showed no evidence of general tonal disadvantages at the level of word recognition. This study extends evidence of a benefit for lexical tone learning from Cooper & Wang (2012)—which looked at naïve tone word learning—to advanced L2 learners. In this way, we capture a fuller picture of how properties of the Mandarin lexicon (tone gaps, disyllabic words) influence L2 tone acquisition in the long-term.

If we had stopped at the level of phonetic category formation, we might well have concluded that there were no clear benefits for L1 tone experience. By extending our investigation to the level of lexical processing and encoding, we gain a fuller picture of cross-linguistic influences, and can see how L1 tone experience has to potential to convey some advantages for tone word learning.

We now turn our attention to these different levels of learning in more detail.

4.1 The influence of L1 tones on the acquisition of L2 tones

Present data is complementary to many other studies in showing unique difficulties for cross-linguistic tone acquisition due to the influences of a speaker's L1 tone categories on their perception of L2 tones. In previous studies, this type of outcome has often been described with reference to models like PAM (Hallé et al., 2004; So & Best, 2010, 2014) or L2LP (cf. Wiener et al., 2019). The present tone identification results do not provide for a strong test of predictions from these models, nevertheless, given the prominence of PAM in the L2 tone literature (e.g.,

Best, 2019; Chen et al., 2020; Hallé et al., 2004; Reid et al., 2015; So & Best, 2010, 2014), we will briefly consider how present results might fit within the PAM framework.

Vietnamese listeners were less accurate than Mandarin listeners for T1 and T4 in all three contexts tested in the present study (MS, DS, CS), with error patterns that suggest they generally confuse these two tones with one another. Perhaps the most telling case was that they confused the low-falling allophone of T3 in CS context with both T4 and T1 (rather than exclusively T4, as Mandarin listeners did). Given the mutual confusability of T1 and T4, present results suggest Vietnamese listeners may perceive both as equally good (or poor) exemplars of Vietnamese native tone categories (so-called Single-Category assimilation). Although a phonetic analysis suggests T4 should be a worse exemplar of L1 Vietnamese categories than T1 (perhaps even Uncategorizable), the bidirectional error pattern argues against an asymmetrical relationship in how these two tones were perceived. On the other hand, learners' survey responses suggest just such an asymmetry as they often described T1, but not T4, as a good fit to a specific Vietnamese tone (*ngang*). Tests of perceptual assimilation and discrimination would be needed to fully interpret these identification patterns.

While not necessarily in conflict with PAM, cue-based accounts (Francis et al., 2008; Francis & Nusbaum, 2002; Holt & Lotto, 2010) might ascribe the apparent confusability of T1 and T4 to the extra weight Vietnamese listeners assign to F0 onset cues (Brunelle, 2009; Li et al., 2017), given that T1 and T4 both have high F0 onsets. This would nicely account for present results, however, it is somewhat surprising that the steep F0 fall of T4 is not more salient, given that Vietnamese participants also give significant weight to F0 change (ΔF0, e.g., Li et al., 2017). Alternatively, Vietnamese performance might be attributed to other cues, such as duration. Especially in isolated MS, T1 often displays a longer duration than T4 (this can be observed

clearly in the tones in Figure 3, though not in Figure 2). One reviewer suggested that durational cues might aid Mandarin listeners, and that an inability to utilize these cues could conceivably account for Vietnamese T1-T4 confusions. For a clear answer, future work will need to conduct targeted tests of specific cues.

While L1 Vietnamese tones seemed to negatively influence perception of T1 and T4, they may have provided some benefit for the perception of T2 and T3. Vietnamese participants consistently performed at nativelike levels of accuracy for T2 and T3 across all contexts, with only T2 in CS showing significantly lower accuracy than the Mandarin group—but without any clear direction to the apparent confusions (errors were fairly evenly dispersed across T1, T3, and T4).

In PAM terms, this strong performance for these tones might be attributed to categorized assimilation of T2 and T3 to L1 Vietnamese tone categories. Based on surveys, the most common pattern was T2 is similar to *sắc* (rising), and T3 to *hỏi* (falling-rising) respectively. Future work could test such patterns more directly.

Cue-based accounts might attribute Vietnamese T2 and T3 identification accuracy to the perceptual weight listeners give to F0 change, and might also posit a role for voice quality cues. Due to its low pitch target, T3 often displays creakiness (though this is not exclusive to T3, cf. Kuang, 2017). The creakiness of T3, then, could potentially aid (Northern) Vietnamese listeners given their use of this cue in their L1 (Brunelle, 2009). The current study did not fully control voice quality, but did attempt to minimize its impact by selecting T3 MS stimuli that were minimally creaky. If listeners (L1 or L2) rely on creakiness to aid in identification of T3 in MS, this would make T3-as-T2 errors more likely in that context, and so may be a partial explanation for the observed T3 MS error patterns of both L1 and L2 groups.

4.2 Contextual influences on tone identification accuracy

The concern for the influence of context in the present study was primarily motivated by conflicting evidence for the 'difficulty' of T3 in past studies, with some studies showing T3 to be relatively difficult even for L1 listeners to identify accurately (Huang & Johnson, 2010), and others suggesting it can be the easiest of all the tones, even for naïve listeners (Chang & Bowles, 2015). We suspect that opposing outcomes like these are due to the idiosyncrasies of specific T3 stimuli used in different studies. For example, when duration is similar (or even normalized) and creakiness is minimal, isolated T2 and T3 stimuli become more similar, and thus more confusable. Alternatively, if some T3 stimuli display longer durations or obvious creakiness, they may be easier to distinguish from T2.

We found that T3 in MS produced in isolation induced frequent confusions with T2 for both L1 and L2 listeners. This is likely due to the phonetic similarity of the dip in both tones (Hao, 2012; Shen & Lin, 1991), and was likely further exacerbated by the variability in F0 height across our four talkers. Talker variability likely reduced listeners' ability to normalize the onset F0 height of T2 and T3. The contrast of MS and DS outcomes for T3 in the present study highlights that such phonetic similarities may apply when T3 is produced in isolation, but rarely do when there is a following syllable. Presenting T3 in DS contexts largely alleviated T3-as-T2 confusions for both Vietnamese and Mandarin listeners.

Whereas T3 was identified with high accuracy in DS, in CS both Vietnamese and Mandarin listeners overwhelmingly confused the clipped T3 as the high-falling T4. Once again, this makes sense given their phonetic similarity, and the variability in onset F0 presented across talkers. This outcome replicates and extends Pelzl (2018), where the same stimuli and task were

used with only L1 Mandarin listeners. It also resembles results reported by C. Han et al. (2020) who tested tone identification by L1 Mandarin listeners using syllables clipped from continuous speech.

Mandarin listeners have been shown to use preceding context as a 'frame of reference' to interpret ambiguous tones (Huang & Holt, 2009; Moore & Jongman, 1997). We suggest both T3 MS and CS error patterns can be understood as reflecting tonal language listeners' sensitivity to the relative height of F0 on the following syllable. Whereas T3 allophones may be ambiguous outside of context, in context they are easily identified. It seems, however, that this same contextual support did not aid Vietnamese listeners in identifying T1 and T4. DS stimuli in the present study used the neutral tone on the second syllable, future work might explore how coarticulation with fully realized tones impacts identification.

4.3 Comparison with L1 English speakers in PLGD21

4.3.1 Identification of Mandarin tone categories

The present study provides a contrast to results from a similar tone identification task (without clipped syllables) reported in Pelzl (2018; with the same participants as described in PLGD21). The inclusion of clipped syllables in the present study and not in PLGD21 should be borne in mind while making comparisons between the two sets of results, as this difference might have had impacts on response patterns.

In contrast to Vietnamese participants who were least accurate for T1 and T4, L1 English participants were highly accurate in identifying T1 (m=99%) and T4 (m=98%). They also achieved 90% accuracy for T2. Like Mandarin and Vietnamese participants in the present study, L1 English participants showed lower accuracy for T3 in MS (m=71%). Unlike those groups, L1

English accuracy decreased for T3 in DS (m=68%), suggesting listeners were unable to capitalize on contextual cues the way Vietnamese and Mandarin listeners were in the present study. L1 English accuracy in DS also decreased for T1 (m=81%) and T2 (m=80%), but not T4 (m=97%). The influences that drive non-tonal patterns of tone identification have long been somewhat of a puzzle (see recent discussion in Best, 2019) and will not be solved here. What is clear, however, is that L1 experience exerts specific and lasting influences on L2 tone perception, well into advanced levels of proficiency.

4.3.2 Recognition of Mandarin tone words

Results of the lexical decision task highlight a striking difference between the performance of L1 Vietnamese and PLGD21's L1 English learners (Figure 8; for statistical analyses and additional visualization, see supplementary materials A4). The Vietnamese group showed no strong difference between vowel and tone nonwords (Vietnamese m=74%; English m=86%). It was less accurate than the English group in rejecting vowel nonwords (Vietnamese m=74%; English m=86%), and descriptively (but not statistically) more accurate for tone nonwords (Vietnamese m=70%; English m=62%). The difference between nonword conditions (23%) was significant for the English group, and more than five times as large as the difference for the Vietnamese group (4%). These patterns persisted for both groups after accounting for vocabulary knowledge in the 'best case scenario' analysis. Individual participant d′ scores show that group results are representative of individual trends. Vietnamese participants varied as to whether they were more sensitive for vowel or tone nonwords. Nineteen Vietnamese participants found tone nonwords more difficult to detect than vowel nonwords; eleven had the opposite pattern, finding vowel nonwords more difficult to detect than tone nonwords; three had equal scores in both. In contrast,

of the eighteen L1 English participants in PLGD21, all but one had higher d′ for vowel compared to tone nonwords.

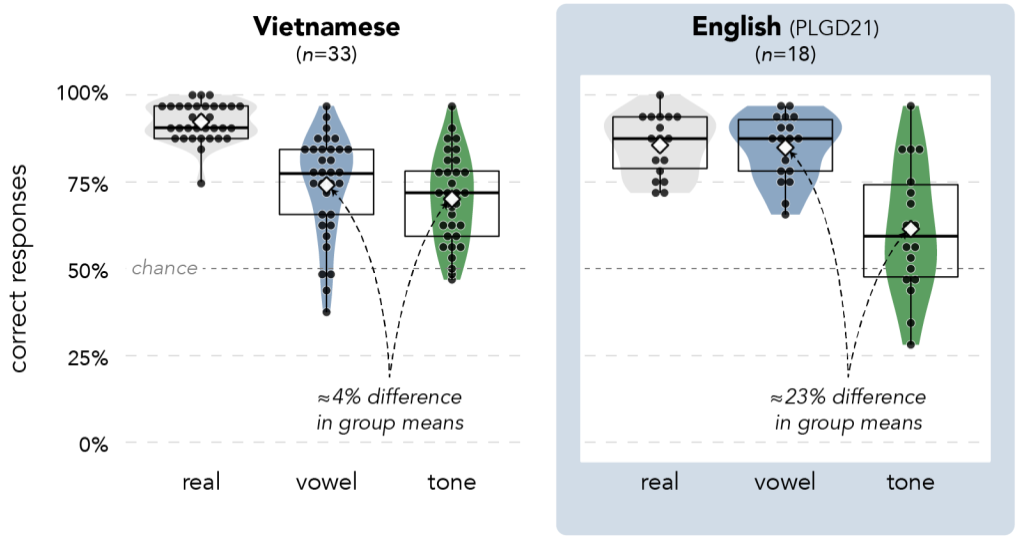

FIGURE 8. Comparison of lexical decision results for L1 Vietnamese participants and L1 English participants (English results adapted from Pelzl et al. 2021).

Still, the pattern of results does not perfectly fit our predictions. The large between-group differences are driven not by the Vietnamese group's superior performance on tone nonwords, but instead by its lower accuracy on vowel nonwords. This allows for at least two plausible interpretations of the data. On the one hand, if we assume that both groups are equally proficient in Mandarin, results could be interpreted as evidence that, compared to English listeners, Vietnamese listeners have greater difficulty using Mandarin vowels for word recognition. However, we know of no reason to expect that Vietnamese learners' Mandarin proficiency would plateau in this way.[4] Moreover, roughly a third of Vietnamese performed with native-like

[4] Vietnamese speakers are not, to our knowledge, known for having difficulty with Mandarin vowels—though there is admittedly little research in this area. In a small qualitative study of L2 vowel production, Gu (2014) reports that novice L1 Vietnamese speakers' production of Mandarin apical vowels (e.g., [ɿ] and [ʅ]) were similar to the high front vowel /i/ in Mandarin, however, the productions of more proficient speakers closely resembled L1 Mandarin vowels. In other words, even if these specific difficulties applied generally to Vietnamese perception of Mandarin vowels, we would not expect them to strongly affect the current sample of more proficient Vietnamese speakers.

sensitivity (d′) for *both* Mandarin vowels and tones (see Figure A2.2 in the supplementary materials), compared to just 3 of 18 English participants in PLGD21. So then, the more likely explanation for the lower average accuracy of the Vietnamese participants is that, as a group, they were less proficient than the English participants in PLGD21. Although we used the same screening tests for both groups, these tests only established a lower bound on proficiency. Average length of study and immersion were still substantially longer for PLGD21's English participants (study length: 8.3 years for English, 5.7 for Vietnamese; length of immersion: 3.5 years for English, 1.6 years for Vietnamese). Assuming the English group was indeed more proficient in Mandarin, results further highlight the continued difficulty English listeners have with tones: even when they are more advanced in L2 Mandarin than tonal L1 listeners, they are still less accurate on tone nonwords.

Beyond the confines of the current study, English speakers are well-known for having difficulty perceiving and producing Mandarin tones, and a growing number of studies have found that a variety of non-tonal L1 speakers at relatively advanced levels of L2 Mandarin proficiency show pronounced difficulties with Mandarin tones in lexical tasks (*L1 English:* Pelzl et al, 2019; Ling & Grüter, 2022; *L1 Korean:* Han & Tsukada, 2020; *L1 Dutch:* Zou et al., 2022). As noted earlier, Cooper and Wang (2012) also found an advantage for naïve listeners from tonal (Cantonese) over non-tonal (English) L1s when learning novel tone words. Given this broader context, we believe present results reflect a real difference in tone word learning between Vietnamese and English speakers.

We posit that tonal and non-tonal language speakers face qualitatively different challenges when acquiring a tonal L2. Non-tonal L1 speakers must establish new tone

More informally, we also consulted with two L1 Vietnamese learners of L2 Mandarin regarding Mandarin vowels. Both indicated that they had not found the vowels to be particularly challenging to learn.

categories—potentially this is an entirely novel *class* of phonological contrasts for these speakers. They must partition F0 cues into discrete categories, and learn how to utilize those categories functionally to differentiate lexical units. To accomplish this, they must overcome their L1 processing biases (Chang, 2018; MacWhinney & Bates, 1989; Strange, 2011). In contrast, tonal L1 speakers are adding new categories to an already existing tone space. This may cause difficulties as the new and previously learned tone categories interact, and learners may need to learn to reweight specific tone cues (F0 height, F0 onset) in order to accurately use the new tones. However, they do not need to newly learn to process tones as functional lexical cues.

Another way to describe these differences is with respect to the *integrality* of a listener's perception of segmental and tonal cues (Lee & Nusbaum, 1993). Depending on their L1 experience, listeners might perceive F0 as a separable (non-lexical) source of information, or as an integrated source of information that works together with co-occurring segments to guide word recognition. Although our Vietnamese participants were overall less accurate than L1 Mandarin participants, their performance could still be described as consistent with integrated perception; they showed no strong bias for tonal vs. segmental information during lexical decision. In contrast, PLGD21's English participants demonstrated a clear bias for using segmental information, indicating that they had not fully integrated tones and segments. Superficially, these results differ from those reported in Zou et al. (2017) who found that non-tonal (L1 Dutch) speakers who had achieved advanced L2 Mandarin proficiency were able to integrate tonal and segmental cues similarly to L1 Mandarin listeners on a nonword ABX task. We suggest a key reason for the apparently different outcomes for Vietnamese and English participants compared to Zou and colleague's Dutch participants is that our task was ultimately more challenging. Among many other differences, our lexical decision task was fully lexical: in

order to successfully reject nonwords, listeners had to compare auditory stimuli to their stored mental representations of those words. In contrast, Zou and colleague's nonword ABX task did not require lexical access. This could explain why lexical decision results show differences from L1 Mandarin listeners while ABX results did not (see also Zou et al., 2022, where differences between L1 Dutch and L1 Mandarin listeners were found using a lexical decision task).

In the end, it may not be particularly helpful to claim that L2 tone acquisition is more or less difficult for tonal language speakers. Depending on the nature of the L1 tone inventory, this process may at times be easier or harder than the challenges facing the non-tonal L2 learner. However, it is important to note the difference in the learning tasks that each group of learners faces. As described already, this bears on theoretical accounts of cross-linguistic tone influence, but it also has practical implications. For instance, it might suggest the use of different pedagogical interventions for non-tonal vs. tonal learners. While both tonal and non-tonal speakers might benefit from teachers paying attention to the defining features of tones (height, direction, onset), L1 tonal learners might receive additional benefit from targeted instruction on those features of the tonal inventories that are poorly aligned and prone to cause confusion. For example, a teacher might tell Vietnamese learners to pay more attention to F0 direction for T4, rather than F0 onset.

One outcome that was—at least to the authors—somewhat surprising was the apparent degree of difficulty Vietnamese participants have in remembering tones for Mandarin words. In this sense, they look quite similar to the English participants in PLGD21. Even when they were confident of their knowledge, they were still incorrect in approximately 15% of the time. While exploratory analyses (see Supplementary Materials Appendix A2) suggest these errors were more common for words containing T1 or T4 on the first syllable, they were nonetheless

common across all tones (range of accuracies: min=60%; max=92%). In other words, while Vietnamese speakers have a general advantage in using tones functionally for word recognition, this does not automatically translate into superior memory for tones in words. While this superficially resembles patterns of tone forgetfulness found in English speakers in PLGD21, it seems likely that—like tone identification results—there are different L1 effects that result in tone forgetfulness for tonal and non-tonal speakers. Future work might probe these issues in more detail.

4.4 Limitations

The main limitation of the present study is the lack of certainty about proficiency matching between Vietnamese participants and the English participants in PLGD21. Although we used the same proficiency measures in the present study as in PLGD21, results suggest that L2 proficiency was not fully matched. Future work might use study and immersion time as additional screening measures to more closely match proficiency between groups.

Other limitations of the present study were 'baked in' by the decision to closely imitate the stimuli and design of PLGD21. While this choice allowed for close comparison between this study and that one, it also meant that we did not address some of the specific characteristics of Vietnamese language that make it an interesting language in relation to Mandarin. For instance, we did not attempt to control for influences of Sino-Vietnamese borrowings on word recognition, nor did we attempt to control the interplay of segmental features between Vietnamese and Mandarin. Because we did not control durational cues, and attempted to minimize voice quality cues, the present study also does not provide strong information about Vietnamese listeners' cue-

weighting. Future work might gain additional insights into cross-linguistic influences by designing stimuli and tasks with such issues in mind.

5.0 CONCLUSION

This study provides new evidence of the cross-linguistic influence of tonal L1 experience on the perception and acquisition of L2 tones. Extending previous research, we found evidence that the influence of L1 tones on L2 tone category formation is persistent into relatively advanced stages of L2 acquisition. Unlike non-tonal L1 speakers in previous studies, L1 Vietnamese speakers with advanced L2 proficiency in Mandarin showed no significant disadvantage for distinguishing real words and nonwords using tones compared to vowels. We interpret this as evidence that tonal L1 speakers are adept at integrating tones into the phonological representations of words—a function they apply naturally due to their extensive L1 *lexical* tone experience.

Acknowledgments

We wish to thank Chu Thị Ngọc Anh and Nguyễn Thị Vân for their help with Vietnamese language translation. This research would not have been possible without their assistance.

Funding sources

This material is based upon work supported by the National Science Foundation under Grant No. 2004279. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Alves, M. J. (2009). Loanwords in Vietnamese. In M. Haspelmath & U. Tadmor (Eds.), *Loanwords in the World's Languages: A Comparative Handbook* (pp. 617–637). De Gruyter Mouton. https://doi.org/10.1515/9783110218442

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious Mixed Models. *ArXiv:1506.04967 [Stat]*. http://arxiv.org/abs/1506.04967

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Best, C. T. (2019). The Diversity of Tone Languages and the Roles of Pitch Variation in Non-tone Languages: Considerations for Tone Perception Research. *Frontiers in Psychology*, *10*, 364. https://doi.org/10.3389/fpsyg.2019.00364

Boersma, P., & Weenink, D. (2018). *Praat: Doing phonetics by computer* (6.0.42). www.praat.org

Bohn, O.-S., & Best, C. T. (2012). Native-language phonetic and phonological influences on perception of American English approximants by Danish and German listeners. *Journal of Phonetics*, *40*(1), 109–128. https://doi.org/10.1016/j.wocn.2011.08.002

Brunelle, M. (2009). Tone perception in Northern and Southern Vietnamese. *Journal of Phonetics*, *37*(1), 79–96. https://doi.org/10.1016/j.wocn.2008.09.003

Chandrasekaran, B., Sampath, P. D., & Wong, P. C. M. (2010). Individual variability in cue-weighting and lexical tone learning. *The Journal of the Acoustical Society of America*, *128*(1), 456–465.

Chang, C. B. (2018). Perceptual attention as the locus of transfer to nonnative speech perception. *Journal of Phonetics*, *68*, 85–102. https://doi.org/10.1016/j.wocn.2018.03.003

Chang, C. B., & Bowles, A. R. (2015). Context effects on second-language learning of tonal contrasts. *Journal of the Acoustical Society of America*, *136*(6), 3703–3716.

Chang, C. B., & Mishler, A. (2012). Evidence for language transfer leading to a perceptual advantage for non-native listeners. *The Journal of the Acoustical Society of America*, *132*(4), 2700–2710. https://doi.org/10.1121/1.4747615

Chang, Y. S., Yao, Y., & Huang, B. H. (2017). Effects of linguistic experience on the perception of high-variability non-native tones. *The Journal of the Acoustical Society of America*, *141*(2), EL120–EL126. https://doi.org/10.1121/1.4976037

Chen, J., Best, C. T., & Antoniou, M. (2020). Native phonological and phonetic influences in perceptual assimilation of monosyllabic Thai lexical tones by Mandarin and Vietnamese listeners. *Journal of Phonetics*, *83*, 101013. https://doi.org/10.1016/j.wocn.2020.101013

Chen, Y., & Xu, Y. (2006). Production of Weak Elements in Speech – Evidence from F_0 Patterns of Neutral Tone in Standard Chinese. *Phonetica*, *63*(1), 47–75. https://doi.org/10.1159/000091406

Cooper, A., & Wang, Y. (2012). The influence of linguistic and musical experience on Cantonese word learning. *The Journal of the Acoustical Society of America*, *131*(6), 4756–4769.

Duanmu, S. (2007). *The Phonology of Standard Chinese* (2nd edition). Oxford University Press.

Francis, A. L., Ciocca, V., Ma, L., & Fenn, K. (2008). Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *Journal of Phonetics*, *36*, 268–294.

Francis, A. L., & Nusbaum, H. C. (2002). Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(2), 349–366. https://doi.org/10.1037//0096-1523.28.2.349

Gandour, J. T. (1983). Tone perception in Far Eastern languages. *Journal of Phonetics*, *11*, 149–175.

Gandour, J. T., & Harshman, R. A. (1978). Crosslanguage Differences in Tone Perception: A Multidimensional Scaling Investigation. *Language and Speech*, *21*(1), 1–33. https://doi.org/10.1177/002383097802100101

Gårding, E., Kratochvil, P., Svantesson, J.-O., & Zhang, J. (1986). Tone 4 and Tone 3 Discrimination in Modern Standard Chinese. *Language and Speech*, *29*(3), 281–293. https://doi.org/10.1177/002383098602900307

Hallé, P. A., Chang, Y.-C., & Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of Phonetics*, *32*, 395–421.

Han, C., Vogel, I., Yuan, Y., & Athanasopoulou, A. (2020). Perceptual Confusion of Mandarin Tone 3 and Tone 4. *University of Pennsylvania Working Papers in Linguistics*, *26*(1).

Han, J.-I., & Tsukada, K. (2020). Lexical representation of Mandarin tones by non-tonal second-language learners. *The Journal of the Acoustical Society of America*, *148*(1), EL46–EL50. https://doi.org/10.1121/10.0001586

Han, M. (1969). *Studies in the phonology of Asian languages VIII: Vietnamese tones.* (Technical Report No. AD0687519). University of Southern California Los Angeles Acoustics Phonetics Research Lab.

Hao, Y.-C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, *40*(2), 269–279. https://doi.org/10.1016/j.wocn.2011.11.001

Holt, L. L., & Lotto, A. J. (2010). Speech perception as categorization. *Attention, Perception & Psychophysics*, *72*(5), 1218–1227. https://doi.org/10.3758/APP.72.5.1218

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal*, *50*(3), 346–363.

Huang, J., & Holt, L. L. (2009). General perceptual contributions to lexical tone normalization. *The Journal of the Acoustical Society of America*, *125*(6), 3983–3994. https://doi.org/10.1121/1.3125342

Huang, T., & Johnson, K. (2010). Language specificity in speech perception: Perception of Mandarin tones by native and nonnative listeners. *Phonetica*, *67*(4), 243–267. https://doi.org/10.1159/000327392

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.

Kuang, J. (2017). Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice. *The Journal of the Acoustical Society of America*, *142*(3), 1693–1706. https://doi.org/10.1121/1.5003649

Lee, L., & Nusbaum, H. C. (1993). Processing interactions between segmental and suprasegmental information in native speakers of English and Mandarin Chinese. *Perception & Psychophysics*, *53*(2), 157–165.

Lee, W.-S., & Zee, E. (2008). Prosodic characteristics of the neutral tone in Beijing Mandarin/北京话轻声的韵律特征. *Journal of Chinese Linguistics*, *36*(1), 1–29.

Lenth, R. (2022). *emmeans: Estimated Marginal Means, aka Least-Squares Means.* (R package version 1.7.4-1). https://CRAN.R-project.org/package=emmeans

Li, B., Shao, J., & Bao, M. (2017). Effects of Phonetic Similarity in the Identification of Mandarin Tones. *Journal of Psycholinguistic Research*, *46*(1), 107–124. https://doi.org/10.1007/s10936-016-9422-6

Ling, W., & Grüter, T. (2022). From sounds to words: The relation between phonological and lexical processing of tone in L2 Mandarin. *Second Language Research*, *38*(2), 289–313. https://doi.org/10.1177/0267658320941546

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide* (2nd ed.). Lawrence Erlbaum Associates.

MacWhinney, B., & Bates, E. (Eds.). (1989). *The Cross-linguistic Study of Sentence Processing*. Cambridge University Press.

Maddieson, I. (2013). Tone. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Max Plank Institute for Evolutionary Anthropology. http://wals.info/chapter/13

Moore, C. B., & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *The Journal of the Acoustical Society of America*, *102*(3), 1864–1877. https://doi.org/10.1121/1.420092

Nguyễn, V. L., & Edmondson, J. (1997). Tones and voice quality in modern northern Vietnamese: Instrumental case studies. *Mon-Khmer Studies*, *28*(35), 1–18.

Nhan, N. T. (1984). *The syllabeme and patterns of word formation in vietnamese* [Dissertation]. New York University.

Pelzl, E. (2018). *Second language lexical representation and processing of Mandarin Chinese tones*. University of Maryland, College Park.

Pelzl, E. (2019). What makes second language perception of Mandarin tones hard?: A non-technical review of evidence from psycholinguistic research. *Chinese as a Second Language (漢語教學研究—美國中文教師學會學報)*, *54*(1), 51–78. https://doi.org/10.1075/csl.18009.pel

Pelzl, E., Lau, E. F., Guo, T., & DeKeyser, R. (2019). Advanced second language learners' perception of lexical tone contrasts. *Studies in Second Language Acquisition*, *41*(1), 59–86. https://doi.org/10.1017/S0272263117000444

Pelzl, E., Lau, E. F., Guo, T., & DeKeyser, R. (2021a). Even in the best-case scenario L2 learners have persistent difficulty perceiving and utilizing tones in Mandarin: Findings from behavioral and event-related potential experiments. *Studies in Second Language Acquisition*, *43*(2), 268–296. https://doi.org/10.1017/S027226312000039X

Pelzl, E., Lau, E. F., Guo, T., & DeKeyser, R. M. (2021b). Advanced Second Language Learners of Mandarin Show Persistent Deficits for Lexical Tone Encoding in Picture-to-Word Form Matching. *Frontiers in Communication*, *6*, 689423. https://doi.org/10.3389/fcomm.2021.689423

Qin, Z., & Jongman, A. (2015). Does second language experience modulate perception of tones in a third language? *Language and Speech*, 0023830915590191.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Reid, A., Burnham, D., Kasisopa, B., Reilly, R., Attina, V., Rattanasone, N. X., & Best, C. T. (2015). Perceptual assimilation of lexical tone: The roles of language experience and visual information. *Attention, Perception, & Psychophysics*, *77*(2), 571–591. https://doi.org/10.3758/s13414-014-0791-3

Schaefer, V., & Darcy, I. (2014). Lexical function of pitch in the first language shapes cross-linguistic perception of Thai tones. *Laboratory Phonology*, *5*(4), 489–522. https://doi.org/10.1515/lp-2014-0016

Shen, X. S., & Lin, M. (1991). A Perceptual Study of Mandarin Tones 2 and 3. *Language and Speech*, *34*(2), 145–156. https://doi.org/10.1177/002383099103400202

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2022). *afex: Analysis of factorial experiments* (R package version 1.1-1). http://cran.r-project.org/package=afex

So, C. K., & Best, C. T. (2010). Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences. *Language and Speech*, *53*(2), 273–293.

So, C. K., & Best, C. T. (2014). Phonetic influences on English and French listeners' assimilation of Mandarin tones to native prosodic categories. *Studies in Second Language Acquisition*, *36*(02), 195–221. https://doi.org/10.1017/S0272263114000047

Strange, W. (2011). Automatic selective perception (ASP) of first and second language speech: A working model. *Journal of Phonetics*, *39*(4), 456–466. https://doi.org/10.1016/j.wocn.2010.09.001

Tsukada, K. (2019). Are Asian Language Speakers Similar or Different? The Perception of Mandarin Lexical Tones by Naïve Listeners from Tonal Language Backgrounds: A

Preliminary Comparison of Thai and Vietnamese Listeners. *Australian Journal of Linguistics*, *39*(3), 329–346. https://doi.org/10.1080/07268602.2019.1620681

Wiener, S., & Goss, S. (2019). Second and third language learners' sensitivity to Japanese pitch accent is additive. *Studies in Second Language Acquisition*, *41*(04), 897–910. https://doi.org/10.1017/S0272263119000068

Wiener, S., Lee, C., & Tao, L. (2019). Statistical Regularities Affect the Perception of Second Language Speech: Evidence From Adult Classroom Learners of Mandarin Chinese. *Language Learning*, *69*(3), 527–558. https://doi.org/10.1111/lang.12342

Wong, P. C. M., & Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics*, *28*(04), 565–585.

Wu, M.-J., & Hu, M.-G. (2004). An analysis of Vietnamese students' error in Chinese Tones. *Shijie Hanyu Jiaoxue (Chinese Teaching in the World)*, *68*(2), 81–87.

Xu, Y. (1997). Contextual tonal variation in Mandarin. *Journal of Phonetics*, *25*, 61–83.

Yip, M. (2002). *Tone*. Cambridge University Press.

Zehr, J., & Schwartz, F. (2018). *PennController for Internet Based Experiments (IBEX)*. https://doi.org/10.17605/OSF.IO/MD832

Zhang, J., & Lai, Y. (2010). Testing the role of phonetic knowledge in Mandarin tone sandhi. *Phonology*, *27*(01), 153–201. https://doi.org/10.1017/S0952675710000060

Zou, T., Caspers, J., & Chen, Y. (2022). Perception of Different Tone Contrasts at Sub-Lexical and Lexical Levels by Dutch Learners of Mandarin Chinese. *Frontiers in Psychology*, *13*, 891756. https://doi.org/10.3389/fpsyg.2022.891756

Zou, T., Chen, Y., & Caspers, J. (2017). The developmental trajectories of attention distribution and segment-tone integration in Dutch learners of Mandarin tones. *Bilingualism:*

Language and Cognition, *20*(5), 1017–1029.

https://doi.org/10.1017/S1366728916000791